FORM PTO-1390 US DEPARTMENT OF COMMERCE REV. 5-93PATENT AND TRADEMARK OFFICE **TRANSMITTAL LETTER TO THE UNITED STATES DESIGNATED/ELECTED OFFICE (DO/EO/US) CONCERNING A FILING UNDER 35 U.S.C. 371**	ATTORNEYS DOCKET NUMBER **P01,0020** U.S. APPLICATION NO. (if known, see 37 CFR 1.5) 09/787698

INTERNATIONAL APPLICATION NO. **PCT/DE99/02846**	INTERNATIONAL FILING DATE **08 SEPTEMBER 1999**	PRIORITY DATE CLAIMED **23 SEPTEMBER 1998**

TITLE OF INVENTION
METHOD AND APPARATUS FOR DETERMINING A SEQUENCE OF ACTIONS FOR A SYSTEM

APPLICANT(S) FOR DO/EO/US
RALF NEUNEIER ET AL.

Applicant herewith submits to the United States Designated/Elected Office (DO/EO/US) the following items and other information:

1. ☒ This is a **FIRST** submission of items concerning a filing under 35 U.S.C. 371.
2. □ This is a **SECOND** or **SUBSEQUENT** submission of items concerning a filing under 35 U.S.C. 371.
3. ☒ This express request to begin national examination procedures (35 U.S.C. 371(f)) at any time rather than delay.
4. ☒ A proper Demand for International Preliminary Examination was made by the 19th month from the earliest claimed priority date.

5. ☒ A copy of International Application as filed (35 U.S.C. 371(c)(2)).
   a. ☒ is transmitted herewith (required only if not transmitted by the International Bureau).
   b. □ has been transmitted by the International Bureau.
   c. □ is not required, as the application was filed in the United States Receiving Office (RO/US)
6. ☒ A translation of the International Application into English (35 U.S.C. 371(c)(2).

7. ☒ Amendments to the claims of the International Application under PCT Article 19 (35 U.S.C. §371(c)(3))
   a. □ are transmitted herewith (required only if not transmitted by the International Bureau).
   b. □ have been transmitted by the International Bureau.
   c. □ have not been made; however, the time limit for making such amendments has NOT expired.
   d. ☒ have not been made and will not be made.

8. □ A translation of the amendments to the claims under PCT Article 19 (35 U.S.C. 371(c)(3)).

9. ☒ An oath or declaration of the inventor(s) (35 U.S.C. 371(c)(4)).

10. □ A translation of the annexes to the International Preliminary Examination Report under PCT Article 36 (35 U.S.C. 371(c)(5)).

Items 11. to 16. below concern other document(s) or information included:

11. ☒ An Information Disclosure Statement under 37 C.F.R. 1.97 and 1.98; **(PTO 1449, Prior Art, Search Report, References)**.

12. ☒ An assignment document for recording. A separate cover sheet in compliance with 37 C.F.R. 3.28 and 3.31 is included.
   (SEE ATTACHED ENVELOPE)

13. ☒ Amendment "A" Prior to Action.
    □ A SECOND or SUBSEQUENT preliminary amendment.

14. ☒ A substitute specification and substitute specification mark-up.

15. ☒ A change of address letter attached to the Declaration.

16. ☒ Other items or information:
   a. ☒ Submission of Drawings

   b. ☒ EXPRESS MAIL #EL 843728583 US dated March 21, 2001.

U.S. APPLICATION NO (if known, see 37 C.F.R. 1.5)	INTERNATIONAL APPLICATION NO	ATTORNEY'S DOCKET NUMBER
09/787698	PCT/DE99/02846	P01,0020

17. ☒ The following fees are submitted:	CALCULATIONS	PTO USE ONLY
BASIC NATIONAL FEE (37 C.F.R. 1.492(a)(1)-(5): Search Report has been prepared by the EPO or JPO $860.00 International preliminary examination fee paid to USPTO (37 C.F.R. 1.482) $690.00 No international preliminary examination fee paid to USPTO (37 C.F.R. 1.482) but international search fee paid to USPTO (37 C.F.R. 1.445(a)(2) $710.00 Neither international preliminary examination fee (37 C.F.R. 1.482) nor international search fee (37 C.F.R. 1.445(a)(2) paid to USPTO $1000.00 International preliminary examination fee paid to USPTO (37 C.F.R. 1.482) and all claims satisfied provisions of PCT Article 33(2)-(4) $ 100.00		
ENTER APPROPRIATE BASIC FEE AMOUNT =	$ 860.00	
Surcharge of $130.00 for furnishing the oath or declaration later than □ 20 □ 30 months from the earliest claimed priority date (37 C.F.R. 1.492(e)).	$	

Claims	Number Filed		Number Extra	Rate		
Total Claims	21	-20 =	1	X $ 18.00	$ 18.00	
Independent Claims	02	- 3 =	0	X $ 80.00	$	
Multiple Dependent Claims				$270.00 +	$	

TOTAL OF ABOVE CALCULATIONS =	$ 878.00	
Reduction by ½ for filing by small entity, if applicable. Verified Small Entity statement must also be filed. (Note 37 C.F.R. 1.9, 1.27, 1.28)	$	
SUBTOTAL =	$ 878.00	
Processing fee of $130.00 for furnishing the English translation later than □ 20 □ 30 months from the earliest claimed priority date (37 CFR 1.492(f)). +	$	
TOTAL NATIONAL FEE =	$ 878.00	
Fee for recording the enclosed assignment (37 C.F.R. 1.21(h). The assignment must be accompanied by an appropriate cover sheet (37 C.F.R. 3.28, 3.31). $40.00 per property +		
TOTAL FEES ENCLOSED =	$ 878.00	
	Amount to be refunded	$
	charged	$

a. ☒ A check in the amount of $878.00 to cover the above fees is enclosed.

b. □ Please charge my Deposit Account No.__________ in the amount of $_______ to cover the above fees. A duplicate copy of this sheet is enclosed.

c. ☒ The Commissioner is hereby authorized to charge any additional fees which may be required, or credit any overpayment to Deposit Account No. 50-1519. **A duplicate copy of this sheet is enclosed.**

NOTE: Where an appropriate time limit under 37 C.F.R. 1.494 or 1.495 has not been met, a petition to revive (37 C.F.R. 1.137(a) or (b)) must be filed and granted to restore the application to pending status.

SEND ALL CORRESPONDENCE TO:

SCHIFF HARDIN & WAITE
PATENT DEPARTMENT
6600 Sears Tower
233 South Wacker Drive
Chicago, Illinois 60606-6473

CUSTOMER NUMBER 26574

Steven H. Noll
SIGNATURE

Steven H. Noll
NAME

28,982
Registration Number

BOX PCT
IN THE UNITED STATES DESIGNATED/ELECTED OFFICE
OF THE UNITED STATES PATENT AND TRADEMARK OFFICE
UNDER THE PATENT COOPERATION TREATY – CHAPTER II

AMENDMENT "A" PRIOR TO ACTION AND SUBMISSION OF SUBSTITUTE SPECIFICATION

APPLICANT(S):	NEUNEIER, R., et al.
ATTORNEY DOCKET NO:	P01,0020
INTERNATIONAL APPLICATION NO:	PCT/DE99/02846
INTERNATIONAL FILING DATE:	8 SEP 1999
INVENTION:	METHOD AND ARRANGEMENT FOR DETERMINING A SEQUENCE OF ACTIONS FOR A SYSTEM

Assistant Commissioner for Patents
Washington, DC 20231

Sir:

Applicants herewith submit an amendment and substitute specification in the captioned PCT application, and respectfully request entry of same prior to examination in the United States National Stage.

IN THE SPECIFICATION

Cancel the specification as filed and insert therefore the substitute specification provided herewith.

IN THE CLAIMS

Cancel claims 1 – 21 as filed and insert therefore new claims 22 - 42 as follows:

- - What is claimed is:

22. (New) A method for computer-aided determination of a sequence of actions for a system having states, the method comprising the steps of:

performing a transition in state between two states on the basis of an action;

determining the sequence of actions to be performed such that a sequence of states results from the sequence of actions;

optimizing the sequence of steps with regard to a prescribed optimization function, including a variable parameter; and

using the variable parameter to set a risk which the resulting sequence of states has with respect to a prescribed state of the system.

23. (New) The method as claimed in claim 22, further comprising the step of:

using approximative dynamic programming for the purpose of determination.

24. (New) The method as claimed in claim 23, further comprising the step of:

basing the approximative dynamic programming Q-learning.

25. (New) The method as claimed in claim 24, further comprising the steps of:

forming an optimization function within Q-learning in accordance with the following rule:

$$OFQ = Q\left(x; w^{a}\right),$$

and

adapting weights of the function approximator in accordance with the following rule:

$$w_{t+1}^{a_t} = w_t^{a_t} + \eta_t \cdot \aleph^{\kappa}\left(d_t\right) \cdot \nabla Q\left(x_t; w_t^{a_t}\right)$$

wherein

$$d_t = r\left(x_t, a_t, x_{t+1}\right) + \gamma \max_{a \in A} Q\left(x_{t+1}, w_t^{a}\right) - Q\left(x_t, w_t^{a_t}\right)$$

26. (New) The method as claimed in claim 23, further comprising the step of:

basing the approximative dynamic programming on TD(λ)-learning.

27. (New) The method as claimed in claim 26, further comprising the steps of:

forming the optimization function within TD(λ)-learning in accordance with the following rule:

OFTD = J(x;w); and

adapting weights of the function approximator are adapted in accordance with the following rule:

$w_{t+1} = w_t + \eta_t \cdot \aleph^{\kappa}(d_t) \cdot z_t$, wherein $d_t = r(w_t, a_t, x_{t+1}) + \gamma J(x_{t+1}; w_t) - J(x_t; w_t)$, $z_t = \lambda \cdot \gamma \cdot z_{t-1} + \nabla J(x_t; w_t)$, and $z_{-1} = 0$.

28. (New) The method as claimed in claim 27, further comprising the step of:

using a technical system to determine the sequence of actions before the determination measured values are measured.

29. (New) The method as claimed in claim 28, further comprising the step of:

subjecting the technical system to open-loop control in accordance with the sequence of actions.

30. (New) The method as claimed in claim 28, further comprising the step of:

subjecting the technical system to closed-loop control in accordance with the sequence of actions.

31. (New) The method as claimed in claim 30, further comprising the step of:

modeling the system as a Markov Decision Problem.

32. (New) The method as claimed in claim 31, further comprising the step of:

using the system in a traffic management system.

33. (New) The method as claimed in claim 31, further comprising the step of:

using the system in a communications system.

34. (New) The method as claimed in claim 31, further comprising the step of:

using the system to carry out access control in a communications network.

35. (New) The method as claimed in claim 31, further comprising the step of:

using the system to carry out routing in a communications network.

36. (New) A system for determining a sequence of actions for a system having states, wherein a transition in state between two states is performed on the basis of an action, the system comprising:

a processor for determining a sequence of actions, whereby a sequence of states resulting from the sequence of actions is optimized with regard to a prescribed optimization function, and the optimization function includes a variable parameter for setting a risk which the resulting sequence of states has with respect to a prescribed state of the system.

37. (New) The system as claimed in claim 36, wherein the processor is used to subject a technical system to open-loop control.

38. (New) The system as claimed in claim 36, wherein the processor is used to subject a technical system to closed-loop control. - -

39. The system as claimed in claim 36, wherein the processor is used in a traffic management system.

40. The system as claimed in claim 36, wherein the processor is used in a communications system.

41. The system as claimed in claim 36, wherein the processor is used to carry out access control in a communications network.

42. The system as claimed in claim 36, wherein the processor is used to carry out routing in a communications network. - -

IN THE ABSTRACT

Cancel the Abstract as filed, and insert therefore on a separate page, the following Abstract of the disclosure:

- - ABSTRACT OF THE DISCLOSURE

A determination of a sequence of actions is performed such that a sequence of states resulting from the sequence of actions is optimized using a prescribed optimization function. The optimization function includes a variable parameter with which it is possible to establish a risk relating to the resulting sequence of states based upon a prescribed state of the system. - -

REMARKS

A substitute specification and a proper abstract of the disclosure are provided herewith which make editorial changes in order to conform to standard US practice. A marked-up copy of the specification is also provided reflecting the changes made.

In addition, the claims as filed have been canceled and replaced by new claims that more clearly set forth the subject matter of Applicants' invention.

No new matter has been inserted into the application.

Applicants submit that this application is in proper condition for examination in the United States National Stage, which action is earnestly solicited.

Respectfully submitted,

Steven H. Noll

Steven H. Noll (Reg. No. 28,982)

SCHIFF HARDIN & WAITE
Patent Department
6600 Sears Tower
233 South Wacker Drive
Chicago, IL 60606
Telephone: (312) 258-5790
Attorneys for Applicant(s)

Customer Number: 26574

BOX PCT

IN THE UNITED STATES DESIGNATED/ELECTED OFFICE
OF THE UNITED STATES PATENT AND TRADEMARK OFFICE
UNDER THE PATENT COOPERATION TREATY – CHAPTER II

SUBMISSION OF DRAWINGS

APPLICANT(S):	NEUNEIER, R., et al.
ATTORNEY DOCKET NO:	P01,0020
INTERNATIONAL APPLICATION NO:	PCT/DE99/02846
INTERNATIONAL FILING DATE:	8 SEP 1999
INVENTION:	METHOD AND ARRANGEMENT FOR DETERMINING A SEQUENCE OF ACTIONS FOR A SYSTEM

Assistant Commissioner for Patents
Washington, DC 20231

Sir:

Applicants herewith submit four drawing sheets, showing Figures 1 – 6, in the captioned PCT application.

Respectfully submitted,

Steven H. Noll

Steven H. Noll (Reg. No. 28,982)

SCHIFF, HARDIN & WAITE
Patent Department
6600 Sears Tower
233 South Wacker Drive
Chicago, IL 60606
Telephone: (312) 258-5790
Attorneys for Applicant(s)

Customer Number: 26574

FIG 1

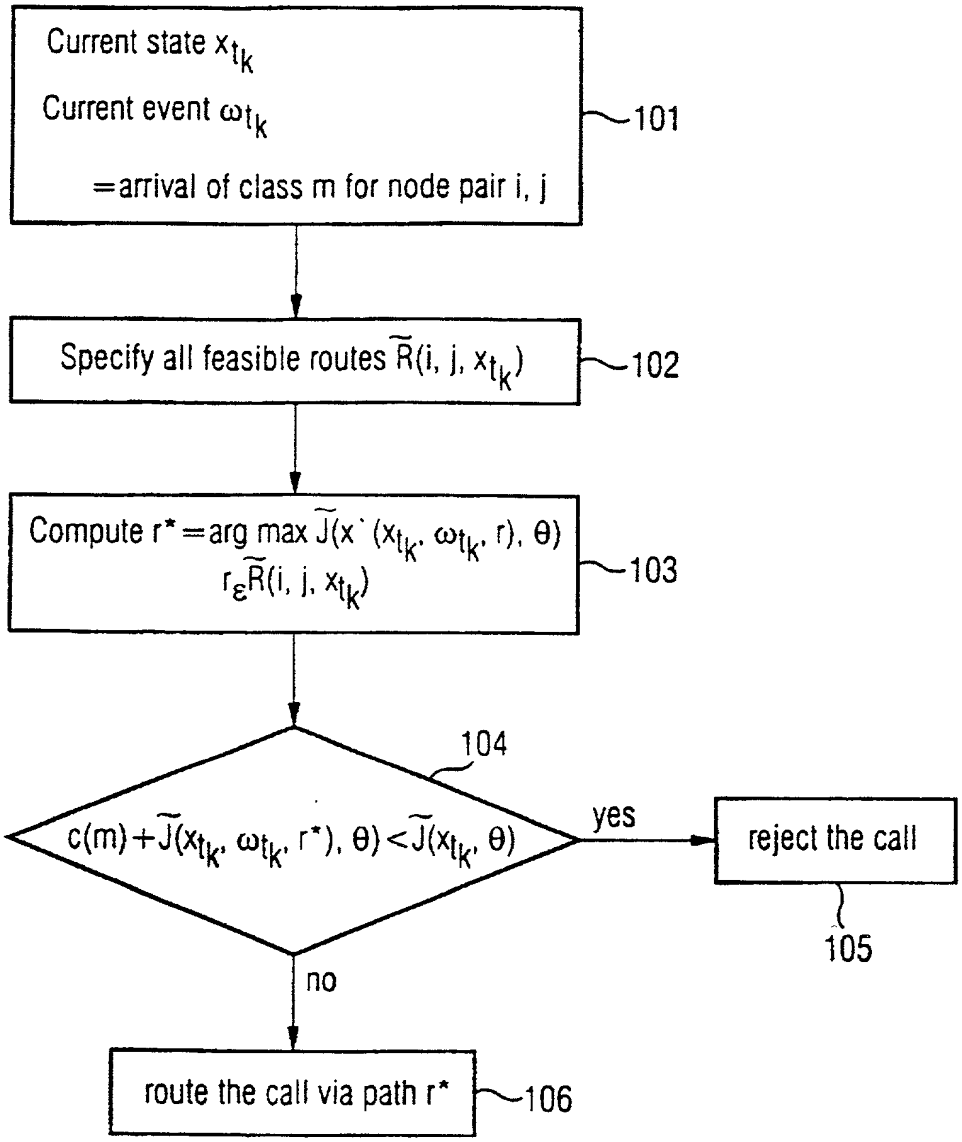

Current state x_{t_k}
Current event ω_{t_k}
$=$arrival of class m for node pair i, j
101
Specify all feasible routes $\widetilde{R}(i, j, x_{t_k})$
102
Compute $r^* = \arg\max_{r \in \widetilde{R}(i, j, x_{t_k})} \widetilde{J}(x^{\prime}(x_{t_k}, \omega_{t_k}, r), \theta)$
103
104
$c(m) + \widetilde{J}(x_{t_k}, \omega_{t_k}, r^*), \theta) < \widetilde{J}(x_{t_k}, \theta)$
yes
reject the call
105
no
route the call via path r*
106

FIG 2

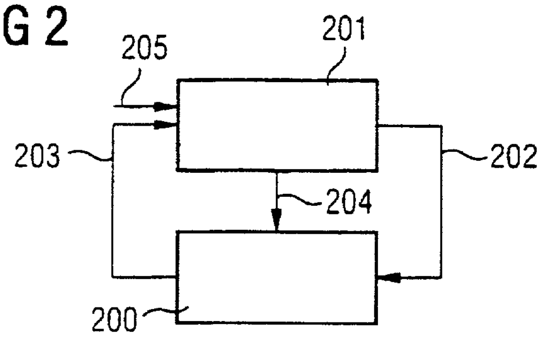
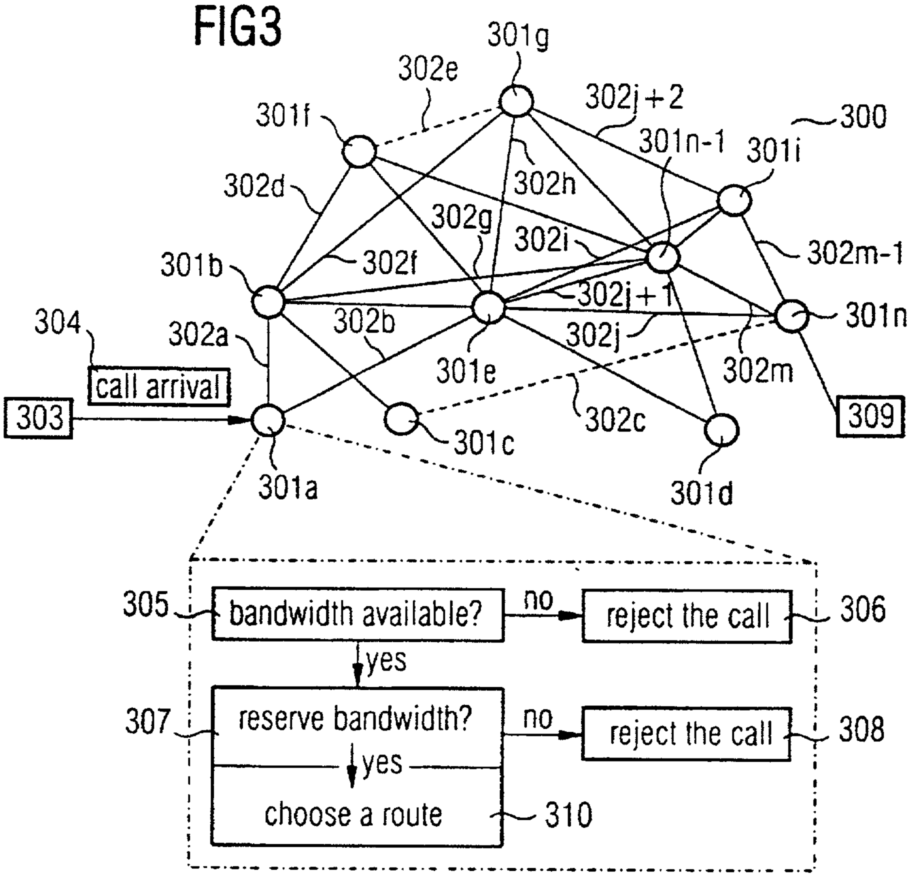

205
201
203
202
204
200

FIG3


301g
302e
302j+2
300
301f
301n-1
301i
302h
302d
302g
302i
302m-1
301b
302f
302j+1
301n
304
302a
302b
302j
call arrival
301e
302m
303
309
301c
302c
301a
301d
305
bandwidth available?
no
reject the call
306
yes
307
reserve bandwidth?
no
reject the call
308
yes
choose a route
310

09/787698

FIG 4

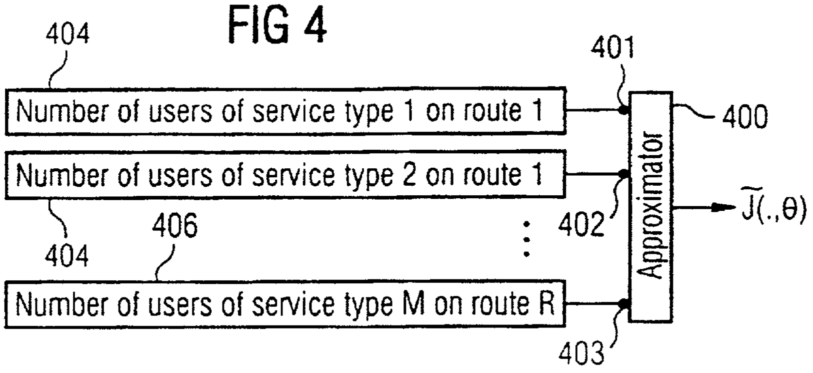

404
401
Number of users of service type 1 on route 1
400
Number of users of service type 2 on route 1
Approximator
$\tilde{J}(.,\theta)$
402
404
406
Number of users of service type M on route R
403

FIG 5

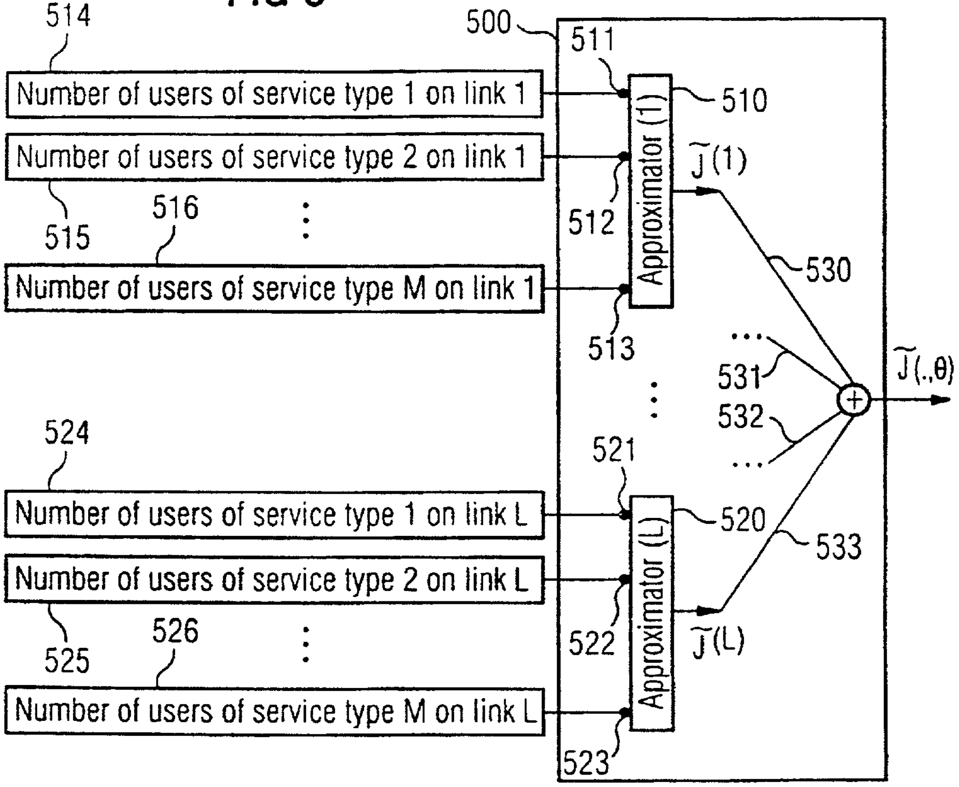
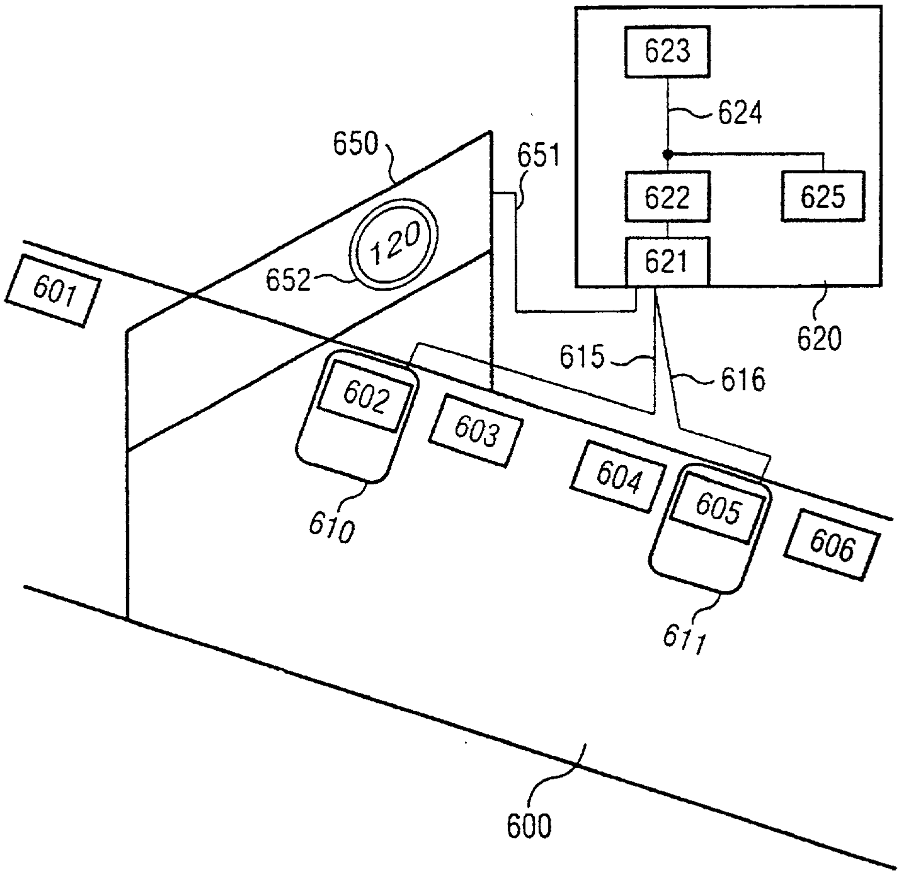

514
500
511
Number of users of service type 1 on link 1
510
Number of users of service type 2 on link 1
Approximator (1)
$\tilde{J}(1)$
512
515
516
530
Number of users of service type M on link 1
513
531
$\tilde{J}(.,\theta)$
532
524
521
Number of users of service type 1 on link L
520
533
Number of users of service type 2 on link L
522
Approximator (L)
$\tilde{J}(L)$
525
526
Number of users of service type M on link L
523

FIG 6


623
624
650
651
622
625
120
601
652
621
620
615
616
602
603
604
605
606
610
611
600

Substitute specification:

- - METHOD AND ARRANGEMENT FOR DETERMINING A SEQUENCE OF ACTIONS FOR A SYSTEM

BACKGROUND OF THE INVENTION

Field of the Invention:

This invention generally pertains to systems having states, and in particular to methods for determining a sequence of actions for such systems.

Discussion of the Related Art:

A generalized method and arrangement for determining a sequence of actions for a system having states, wherein a transition in state between two states is performed on the basis of an action, is discussed by Neuneier in "Enhancing Q-Learning for Optimal Asset Allocation", appearing in the Proceedings of the Neural Information Processing Systems, NIPS 1997. Neuneier describes a financial market as an example of a system which has states. His system is described as a Markov Decision Problem (MDP).

The characteristics of a Markov Decision Problem are represented below by way of summary:

X set of possible states of the system, e.g. $X = \Re^m$,

$A(x_t)$ set of possible actions in the state

$p(x_{t+1} \mid x_t, a_t)$	x_t
$r(x_t, a_t, x_{t+1})$	gain with expectation $R(x_t, a_t)$.

Starting from observable variables, the variables denoted below as training data, the aim is to determine a strategy, that is to say a sequence of functions

$$\pi = \{\mu_0, \mu_1, K\ , \mu_T\}, \tag{3}$$

which at each instant t map each state into an action rule, that is to say action

$$\mu_t(x_t) = a_t \tag{4}$$

Such a strategy is evaluated by an optimization function.

The optimization function specifies the expectation, the gains accumulated over time at a given strategy π, and a start state x_0.

The so-called Q-learning method is described by Neuneier as an example of a method of approximative dynamic programming.

An optimum evaluation function $V^*(x)$ is defined by

$$V^*(x) = \max_{\pi} V^{\pi}(x) \qquad \forall x \in X \tag{5}$$

where

$$V^{\pi}(x) = E\left[\sum_{t=0}^{\infty} \gamma^t r(x_t, \mu_t, x_{t+1}) \middle| x_0 = x\right], \tag{6}$$

γ denoting a prescribable reduction factor which is formed in accordance with the following rule:

$$\gamma = \frac{1}{1+z}, \tag{7}$$

$$z \in \Re^{+}. \tag{8}$$

A Q-evaluation function $Q^*(x_t,a_t)$ is formed within the Q-learning method for each pair (state x_t, action a_t) in accordance with the following rule:

$$Q^*(x_t, a_t) := \sum_{x \in X} p(x_{t+1} | x_t, a_t) \cdot r_t + \\ +\gamma \cdot \sum_{x \in X} p(x | x_t, a_t) \cdot \max_{a \in A}\left(Q^*(x, a)\right) \tag{9}$$

On the basis respectively of the tupel (x_t, x_{t+1}, a_t, r_t), the Q-values Q^* (x,a) are adapted in the k+1 th iteration in accordance with the following learning rule with a prescribed learning rate η_k in accordance with the following rule:

$$Q_{k+1}(x_t, a_t) = (1 - \eta_k)Q_k(x_t, a_t) + \eta_k\left(r_t + \gamma \max_{a \in A}\left(Q_k(x_{t+1}, a)\right)\right). \tag{10}$$

Usually, the so-called Q-values $Q^*(x,a)$ are approximated for various actions by a function approximator in each case, for example a neural network or a polynomial classifier, with a weighting vector w^a, which contains weights of the function approximator.

A function approximator is, for example, a neural network, a polynomial classifier or a combination of a neural network with a polynomial classifier.

It therefore holds that:

$$Q^*(x, a) \approx Q\left(x; w^a\right). \tag{11}$$

Changes in the weights in the weighting vector w^a are based on a temporal difference d_t which is formed in accordance with the following rule:

$$d_t := r\left(x_t, a_t, x_{t+1}\right) + \gamma \max_{a \in A} Q\left(x_{t+1}; w_k^a\right) - Q\left(x_t; w_k^{a_t}\right) \tag{12}$$

The following adaptation rule for the weights of the neural network, which are included in the weighting vector w^a, follows for the Q-learning method with the use of a neural network:

$$w_{k+1}^{a_t} = w_k^{a_t} + \eta_k \cdot d_t \cdot \nabla Q\left(x_t; w_k^{a_t}\right). \tag{13}$$

The neural network representing the system of a financial market as described by Neuneier is trained using the training data which describe information on changes in prices on a financial market as time series values.

A further method of approximative dynamic programming is the so-called TD(λ) learning method. This method is discussed in R.S. Sutton's, "Learning To Predict By The Method Of Temporal Differences", appearing in Machine Learning, Chapter 3, pages 9 - 44, 1988.

Furthermore, it is known from M. Heger's, "Risk and Reinforcement Learning:

Concepts and Dynamic Programming", ZKW Bericht No. 8/94, Zentrum für Kognitionswissenschaften [Center for Cognitive Sciences], Bremen University, December 1994, that risk is associated with a strategy π and an initial state x_t. A method for risk avoidance is also discussed by Hager, cited above.

The following optimization function, which is also referred to as an expanded Q-function $\underline{Q}^{\pi}(x_t, a_t)$, is used in the Hager method:

$$\text{maximize}\left(\underline{Q}^{\pi}(x_t, a_t) := r(x_t, a_t, x_{t+1}) + \inf_{\substack{x_0, x_1, K \\ p(x_0, x_1, K) > 0}} \left\{\sum_{k=1}^{\infty} \gamma^k r(x_k, \pi(x_k), x_{k+1})\right\}\right) \tag{14}$$

The expanded Q-function $\underline{Q}^{\pi}(x_t, a_t)$ describes the worst case if the action a_t is executed in the state x_t and the strategy π is followed thereupon.

The optimization function $\underline{Q}^{\pi}(x_t, a_t)$ for

$$\underline{Q}^{*}(x_t, a_t) := \max_{\pi \in \Pi} \underline{Q}^{\pi}(x_t, a_t) \tag{15}$$

is given by the following rule:

$$\underline{Q}^{*}(x_t, a_t) = \min_{\substack{x \in X \\ p(x_{t+1} | x_t, a_t) > 0}} \left(r(x_t, a_t, x) + \gamma \cdot \max_{a \in A} \underline{Q}^{*}(x, a)\right). \tag{16}$$

A substantial disadvantage of this mode of procedure is that only the worst case is taken into account when finding the strategy. However, this inadequately reflects the requirements of the most varied technical systems.

In "Dynamic Programming and Optimal Control", Athena Scientific, Belmont, MA, 1995, D.P. Bertsekas formulates access control for a communications network and routing within the communications network as a problem of dynamic programming.

Therefore, the present invention is based on the problem of specifying a method and system for determining a sequence of actions in which the method or sequences of actions achieve an increased flexibility in determining the strategy needed.

In a method for computer-aided determination of a sequence of actions for a system which has states, a transition in state between two states being performed on the basis of an action, the determination of the sequence of actions is performed in such a way that a sequence of states resulting from the sequence of actions is optimized with regard to a prescribed optimization function, the optimization function including a variable parameter with the aid of which it is possible to set a risk which the resulting sequence of states has with respect to a prescribed state of the system.

A system for determining a sequence of actions for a system which has states, a transition in state between two states being performed on the basis of an action, has a processor which is set up in such a way that the determination of the sequence of actions can be performed in such a way that a sequence of states resulting from the sequence of actions is optimized with regard to a prescribed optimization function, the optimization function including a variable parameter with the aid of which it is possible to set a risk which the resulting sequence of states has with respect to a prescribed state of the system.

Thus, the present invention offers a method for determining a sequence of actions at a freely prescribable level of accuracy when finding a strategy for a possible closed-loop control or open-loop control of the system, in general for influencing it. Hence, the embodiments described below are valid both for the method and for the system.

Approximative dynamic programming is used for the purpose of determination, for example a method based on Q-learning or a method based on TD(λ)-learning.

Within Q-learning, the optimization function OFQ is preferably formed in accordance with the following rule:

$$OFQ = Q\left(x; w^{a}\right),$$

- x denoting a state in a state space X
- a denoting an action from an action space A, and
- w^{a} denoting the weights of a function approximator which belong to the action a.

The following adaptation step is executed during Q-learning in order to determine the optimum weights w^a of the function approximator:

$$w_{t+1}^{a_t} = w_t^{a_t} + \eta_t \cdot \aleph^{\kappa}\left(d_t\right) \cdot \nabla Q\left(x_t ; w_t^{a_t}\right)$$

with the abbreviation

$$d_t = r\left(x_t, a_t, x_{t+1}\right) + \gamma \max_{a \in A} Q\left(x_{t+1}, w_t^{a}\right) - Q\left(x_t, w_t^{a_t}\right)$$

- x_t, x_t+1 respectively denoting a state in the state space X,
- a_t denoting an action from an action space A,
- γ denoting a prescribable reduction factor,
- $w_t^{a_t}$ denoting the weighting vector associated with the action a_t before the adaptation step,
- $w_{t+1}^{a_t}$ denoting the weighing vector associated with the action a_t after the adaptation step,
- η_t (t = 1, ...) denoting a prescribable step size sequence,
- $\kappa \in [-1; 1]$ denoting a risk monitoring parameter,
- $\aleph^{\kappa}$ denoting a risk monitoring function $\aleph^{\kappa}(\xi) = (1 - \kappa \operatorname{sign}(\xi))\xi$,
- $\nabla Q(;)$ denoting the derivation of the function approximator according to its weights, and
- $r(x_t, a_t, x_{t+1})$ denoting a gain upon the transition of state from the state x_t to the subsequent state x_{t+1}.

The optimization function is preferably formed in accordance with the following rule within the TD(λ)-learning method:

OFTD = $J(x;w)$

- x denoting a state in a state space X,
- a denoting an action from an action space A, and
- w denoting the weights of a function approximator.

The following adaptation step is executed during TD(λ)-learning in order to determine the optimum weights w of the function approximator:

$w_{t+1} = w_t + \eta_t \cdot \aleph^{\kappa}(d_t) \cdot z_t$

with the abbreviations

$d_t = r(w_t, a_t, x_{t+1}) + \gamma J(x_{t+1}; w_t) - J(x_t; w_t)$,

$z_t = \lambda \cdot \gamma \cdot z_{t-1} + \nabla J(x_t; w_t)$,

$z_{-1} = 0$

- x_t, x_{t+1} respectively denoting a state in the state space X,
- a_t denoting an action from an action space A,
- γ denoting a prescribable reduction factor,
- w_t denoting the weighting vector before the adaptation step,
- w_{t+1} denoting the weighting vector after the adaptation step,
- η_t $(t = 1, \ldots)$ denoting a prescribable step size sequence,
- $\kappa \in [-1; 1]$ denoting a risk monitoring parameter,
- $\aleph^{\kappa}$ denoting a risk monitoring function $\aleph^{\kappa}(\xi) = (1 - \kappa \mathrm{sign}(\xi))\xi$,
- $\nabla J(\, ; \,)$ denoting the derivation of the function approximator according to its

weights, and

- $r(x_t, a_t, x_{t+1})$ denoting a gain upon the transition of state from the state x_t to the subsequent state x_{t+1}.

SUMMARY OF THE INVENTION

It is an object of the present invention to provide a technical system and method for determining a sequence of actions using measured values.

It is another object of the present invention to provide a technical system and method that can be subjected to open-loop control or closed-loop control with the use of a determined sequence of actions.

It is a further object of the invention to provide a technical system and method modeled as a Markov Decision Problem.

It is an additional object of the invention to provide a technical system and method that can be used in a traffic management system.

It is yet another object of the invention to provide a technical system and method that can be used in a communications system, such that a sequence of actions is used to carry out access control, routing or path allocation.

It is yet a further object of the invention to provide a technical system and method for a financial market modeled by a Markov Decision Problem, wherein a change in an index of stocks, or a change in a rate of exchange on a foreign exchange market, makes it possible to intervene in the market in accordance with a sequence of determined actions.

These and other objects of the invention will be apparent from a careful review of the following detailed description of the preferred embodiments, which is to read in conjunction with a review of the accompanying drawing figures.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 shows a flowchart of method steps according to the present invention;

Figure 2 shows a system modeled as a Markov Decision Problem;

Figure 3 shows a communications network wherein access control is carried out in a switching unit according to the present invention;

Figure 4 shows a function approximator for approximative dynamic programming according to the present invention;

Figure 5 shows a plurality of function approximators for approximative dynamic programming according to the present invention; and

Figure 6 shows a traffic management system subjected to closed-loop control in accordance with the present invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

Figure 1 shows a flowchart according to the present invention, in which individual method steps of a first embodiment are provided, which will be discussed later.

Figure 2 shows the structure of a typical Markov Decision Problem method.

The system 201 is in a state x_t at an instant t. The state x_t can be observed by an observer of the system. On the basis of an action a_t from a set in the state x_t of possible actions, $a_t \in A(x_t)$, the system makes a transition with a certain probability into a subsequent state x_t+1 at a subsequent instant t+1.

As illustrated diagrammatically in Figure 2 by a loop, an observer 200 perceives 202 observable variables concerning the state x_t and takes a decision via an action 203 with which it acts on the system 201. The system 201 is usually subject to the interference 205.

The observer 200 obtains a gain r_t 204

$$r_t = r(x_t, a_t, x_{t+1}) \in \Re, \qquad (1)$$

which is a function of the action a_t 203 and the original state x_t at the instant t as well as of the subsequent state x_t+1 of the system at the subsequent instant t+1.

The gain r_t can assume a positive or negative scalar value depending on whether the decision leads, with regard to a prescribable criterion, to a positive or negative system development, to an increase in capital stock or to a loss.

In a further time step, the observer 200 of the system 201 decides on the basis of the observable variables 202, 204 of the subsequent state x_{t+1} in favor of a new action a_{t+1}, etc.

A sequence of

State:	x_t	$\in$	X
Action:	a_t	$\in$	$A(x_t)$
Subsequent state:	x_t+1	$\in$	X
Gain	$r_t = r(x_t, a_t, x_{t+1})$	$\in$	$\Re$

describes a trajectory of the system which is evaluated by a performance criterion which accumulates the individual gains r_t over the instants t. It is assumed by way of simplification in a Markov Decision Problem that the state x_t and the action a_t all contain information for the purpose of describing a transition probability $p(x_{t+1}|\cdot)$ of the system from the state x_t to the subsequent state x_{t+1}.

In formal terms, this means that:

$$p(x_{t+1}|x_t, K, x_0, a_t, K, a_0) = p(x_{t+1}|x_t, a_t). \qquad (2)$$

$p(x_{t+1}|x_t,a_t)$ denotes a transition probability for the subsequent state x_{t+1} for a given state x_t and given action a_t.

In a Markov Decision Problem, future states of the system 201 are thus not a function of states and actions which lie further in the past than one time step.

Figure 3 shows an embodiment of the present invention involving an access control and routing system, such as a communications network 300.

The communications network 300 has a multiplicity of switching units 301a, 301b, ..., 301i, ... 301n, which are interconnected via connections 302a, 302b, 302j, ... 302m. A first terminal 303 is connected to a first switching unit 301a. From the first terminal 303, the first switching unit 301a is sent a request message 304 which requests preservation of a prescribed bandwidth within the communications network 300 for the purpose of transmitting data, such as video data or text data.

It is determined in the first switching unit 301a in accordance with a strategy described below, whether the requested bandwidth is available in the communications network 300 on a specified, requested connection instep 305. The request is refused instep 306 if this is not the case. If sufficient bandwidth is available, it is checked in checking step 307 whether the bandwidth can be reserved.

The request is refused in step 308 if this is not the case. Otherwise, the first switching unit 301a selects a route from the first switching unit 301a via further switching units 301i to a second terminal 309 with which the first terminal 303 wishes to communicate, and a connection is initialized in step 310.

The starting point below is a communications network 300 which comprises a set of switching units

$$N = \{1, K\ , n, K\ , N\} \qquad (17)$$

and a set of physical connections

$$\mathcal{L}=\{1,K\ ,l,K\ ,L\},\qquad (18)$$

a physical connection l having a capacity of B(l) bandwidth units.

A set

$$\mathcal{M}=\{1,K\ ,m,K\ ,M\}\qquad (19)$$

of different types of service m are available, a type of service m being characterized by

- a bandwidth requirement b(m),
- an average connection time $\frac{1}{V(m)}$, and
- a gain c(m) which is obtained whenever a call request of the corresponding type of service m is accepted.

The gain c(m) is given by the amount of money which a network operator of the communications network 300 bills a subscriber for a connection of the type of service. Clearly, the gain c(m) reflects different priorities, which can be prescribed by the network operator and which he associates with different services.

A physical connection 1 can simultaneously provide any desired combination of communications connections as long as the bandwidth used for the communications connections does not exceed the bandwidth available overall for the physical connection.

If a new communications connection of type m is requested between a first node i and a second node j (terminals are also denoted as nodes), the requested communications connection can, as represented above, either be accepted or be refused. If the communications connection is accepted, a route is selected from a set of prescribed routes. This selection is denoted as a routing. b(m) bandwidth units are used in the communications connection of type m for each physical connection along the selected route for the duration of the connection.

Thus, during access control, also referred to as call admission control, a route can be selected within the communications network 300 only when the selected route has sufficient bandwidth available. The aim of the access control and of the routing is to maximize a long term gain which is obtained by acceptance of the requested connections.

At an instant t, the technical system which is the communications network 300 is in a state x_t which is described by a list of routes via existing connections, by means of which lists it is shown how many connections of which type of service are using the respective routes at the instant t.

Events w, by means of which a state x_t could be transferred into a subsequent state x_{t+1}, are the arrival of new connection request messages, or else the termination of a connection existing in the communications network 300.

In this embodiment, an action a_t at an instant t, owing to a connection request is the decision as to whether a connection request is to be accepted or refused and, if the connection is accepted, the selection of the route through the communications network 300.

The aim is to determine a sequence of actions, that is to say clearly to determine the learning of a strategy with actions relating to a state x_t in such a way that the following rule is maximized:

$$E\left(\sum_{k=0}^{\infty} e^{-\beta t_k} \cdot g\left(x_{t_k}, \omega_k, a_{t_k}\right)\right), \qquad (20)$$

- E{.} denoting an expectation,
- t_k denoting an instant at which a kth event takes place,
- $g\left(x_{t_k}, \omega_k, a_{t_k}\right)$. denoting the gain which is associated with the kth event, and
- β denoting a reduction factor which evaluates an immediate gain as being more valuable than a gain at instants lying further in the future.

Different implementations of a strategy lead normally to different overall gains G:

$$G = \sum_{k=0}^{\infty} e^{-\beta t_k} \cdot g\left(x_{t_k}, \omega_k, a_{t_k}\right). \qquad (21)$$

The aim is to maximize the expectation of the overall gain G in accordance with the following rule J:

$$J = E\left\{\sum_{k=0}^{\infty} e^{-\beta t_k} \cdot g\left(x_{t_k}, \omega_k, a_{t_k}\right)\right\}, \qquad (22)$$

it being possible to set a risk which reduces the overall gain G of a specific implementation of access control and of a routing strategy to below the expectation.

The TD(λ)-learning method is used to carry out the access control and the routing.

The following target function is used in this embodiment:

$$J^*(x_t) = E_\tau\left\{e^{-\beta\tau}\right\}E_\omega\left\{\max_{a \in A}\left[g(x_t, \omega_t, a) + J^*(x_{t+1})\right]\right\}, \tag{23}$$

- A denoting an action space with a prescribed number of actions which are respectively available in a state x_t,
- τ denoting a first instant at which a first event ω occurs, and
- x_{t+1} denoting a subsequent state of the system.

An approximated value of the target value $J^*(x_t)$ is learned and stored by employing a function approximator 400 (compare Figure 4) with the use of training data.

Training data are data previously measured in the communications network 300 and relating to the behavior of the communications network 300 in the case of incoming connection requests 304 and of termination of messages. This time sequence of states is stored, and these training data are used to train the function approximator 400 in accordance with the learning method described below.

A number of connections of in each case one type of service m on a route of the communications network 300 serve in each case as input variable of the function approximator 400 for each input 401, 402, 403 of the function approximator 400. These are represented in Figure 4 by blocks 404, 405, 406. An approximated target value $\tilde{J}$ of the target value J^* is the output variable of the function approximator 400.

Figure 5 shows a detailed representation of a function approximator 500, which has several component function approximators 510, 520.

One output variable is the approximated target value $\tilde{J}$, which is formed in accordance with the following rule:

$$\tilde{J}(x_t, \Theta) = \sum_{l=1}^{L} \tilde{J}^{(l)}\left(x_t^{(l)}, \Theta_t^{(l)}\right). \tag{24}$$

The input variables of the component function approximators 510, 520, which are present at the inputs 511, 512, 513 of the first component function approximator 510, or at the inputs 521, 522 and 523 of the second component function approximator 520 are, in turn, respectively a number of types of service of a type m in a physical connection r in each case, symbolized by blocks 514, 515, 516 for the first component function approximator, and 524, 525 and 526 for the second component function approximator 520.

Component output variables 530, 531, 532, 533 are fed to an adder unit 540, and the approximated target variable $\tilde{J}$ is formed as output variable of the adder unit.

Let it be assumed that the communications network 300 is in the state x_{tk} and that a request message with which a type of service m of class m is requested for a connection between two nodes i, j reaches the first switching unit 301a.

A list of permitted routes between the nodes i and j is denoted by R(i, j), and a list of all possible routes is denoted by

$$\tilde{R}\left(i, j, x_{t_k}\right) \subset R(i, j) \tag{25}$$

as a subset of the routes R(i, j) which could implement a possible connection with regard to the available and requested bandwidth.

For each possible route r, $r \in \tilde{R}\left(i, j, x_{t_k}\right)$, a subsequent state $x_{t_k} + 1\left(x_{t_k},\ \omega_k, r\right)$ is determined which results from the fact that the connection request 304 is accepted and the connection on the route r is made available to the requesting first terminal 303.

This is illustrated in Figure 1 as step 102, the state of the system and the respective event being respectively determined in step 101. A route r* to be selected is determined in step 103 in accordance with the following rule:

$$r^{*} = \arg \max_{r \in \tilde{R}\left(i, j, x_{t_k}\right)} \tilde{J}\left(x_{t_k} + 1\left(x_{t_k}, \omega_k, r\right), \Theta_t\right). \tag{26}$$

A check is made in step 104 as to whether the following rule is fulfilled:

$$c(m) + \tilde{J}\left(x_{t_k} + 1\left(x_{t_k}, \omega_k, r^{*}\right), \Theta_t\right) < \tilde{J}\left(x_{t_k}, \Theta_t\right). \tag{27}$$

If this is the case, the connection request 304 is rejected in step 105, otherwise the connection is accepted and “switched through” to the node j along the selected route r* in step 106.

Weights of the function approximator 400, 500 which are adapted in the TD(λ)-learning method to the training data, are stored in a parameter vector θ for an instant t in each case, such that an optimized access control and an optimized routing are achieved.

During the training phase, the weighting parameters are adapted to the training data applied to the function approximator.

A risk parameter κ is defined with the aid of which a desired risk, which the system has with regard to a prescribed state owing to a sequence of actions and states, can be set in accordance with the following rules:

$-1 \leq \kappa < 0$:	risky learning,
$\kappa = 0$:	neutral learning with regard to the risk,
$0 < \kappa < 1$:	risk-avoiding learning,
$\kappa = 1$:	worst-case learning.

Furthermore, a prescribable parameter $0 \leq \lambda \leq 1$ and a step size sequence γ_k are prescribed in the learning method.

The weighting values of the weighting vector Θ are adapted to the training data on the basis of each event ω_{t_k} in accordance with the following adaptation rule:

$$\Theta_k = \Theta_{k-1} + \gamma_k \aleph^{\kappa}(d_k) z_t , \qquad (28)$$

in which case

$$d_k = e^{-\beta(t_k - t_{k-1})}\left(g\left(x_{t_k}, \omega_k, a_{t_k}\right) + \tilde{J}\left(x_{t_k}, \Theta_{k-1}\right)\right) - \tilde{J}\left(x_{t_{k-1}}, \Theta_{k-1}\right) \tag{29}$$

$$z_t = \lambda e^{-\beta(t_{k-1} - t_{k-2})} z_{t-1} + \nabla_{\Theta}\tilde{J}\left(x_{t_{k-1}}, \Theta_{k-1}\right), \tag{30}$$

and

$$\aleph^{\kappa}(\xi) = \left(1 - \kappa \mathrm{sign}(\xi)\right)\xi\ . \tag{31}$$

It is assumed that: $z_{-1} = 0$.

The function

$$g\left(x_{t_k}, \omega_k, a_{t_k}\right) \tag{32}$$

denotes the immediate gain in accordance with the following rule:

$$g\left(x_{t_k}, \omega_k, a_{t_k}\right) = \begin{cases} c(m) & \text{when } \omega_{t_k} \text{ is a service request for a type of service } m\text{, and the connection is accepted} \\ 0 & \text{otherwise} \end{cases} \tag{33}$$

Thus, as described above, a sequence of actions is determined with regard to a connection request such that a connection request is either rejected or accepted on the basis of an action. The determination is performed taking account of an optimization function in which the risk can be set by means of a risk control parameter $\kappa \in [-1; 1]$ in a variable fashion.

Figure 6 shows an embodiment of the present invention in relation to a traffic management system

A road 600 on which automobiles 601, 602, 603, 604, 605 and 606 are being driven. Conductor loops 610, 611 integrated into the road 600 receive electric signals in a known way and feed the electric signals 615, 616 to a computer 620 via an input/output interface 621. In an analog-to-digital converter 622 connected to the input/output interface 621, the electric signals are digitized into a time series and stored in a memory 623, which is connected by a bus 624 to the analog-to-digital converter 622 and a processor 625. Via the input/output interface 621, a traffic management system 650 is fed control signals 651 from which it is possible to set a prescribed speed stipulation 652 in the traffic management system 650, or else further particulars of traffic regulations, which are displayed via the traffic management system 650 to drivers of the vehicles 601, 602, 603, 604, 605 and 606.

The following local state variables are used in this case for the purpose of traffic modeling:

- traffic flow rate v,
- vehicle density ρ (ρ = number of vehicles per kilometer $\frac{\mathrm{Fz}}{\mathrm{km}}$).
- traffic flow q (q = number of vehicles per hour $\frac{\mathrm{Fz}}{\mathrm{h}}$, (q= v * ρ)), and
- speed restrictions 652 displayed by the traffic management system 650 at an instant in each case.

The local state variables are measured as described above by using the conductor loops 610, 611.

These variables ($v(t)$, $\rho(t)$, $q(t)$) therefore represent a state of the technical system of "traffic" at a specific instant t.

In this embodiment, the system is therefore a traffic system which is controlled by using the traffic management system 650, and an extended Q-learning method is described as method of approximative dynamic programming.

The state x_t is described by a state vector

$$x(t) = \left(v(t), \rho(t), q(t)\right) . \qquad (34)$$

The action a_t denotes the speed restriction 652, which is displayed at the instant t by the traffic management system 650. The gain $r(x_t, a_t, x_{t+1})$ describes the quality of the traffic flow which was measured between the instants t and t+1 by the conductor loops 610 and 611.

In this embodiment, $r(x_t, a_t, x_{t+1})$ denotes

- the average speed of the vehicles in the time interval [t, t + 1]

or

- the number of vehicles which have passed the conductor loops 610 and 611 in the time interval [t, t + 1]

or

- the variance of the vehicle speeds in the time interval [t, t + 1],

or

- a weighted sum from the above variables.

A value of the optimization function OFQ is determined for each possible action a_t, that is to say for each speed restriction which can be displayed by the traffic management system 650, an estimated value of the optimization function OFQ being realized in each case as a neural network.

This results in a set of evaluation variables for the various actions a_t in the system state x_t. Those actions a_t for which the maximum evaluation variable OFQ has been determined in the current system state x_t are selected in a control phase from the possible actions a_t, that is to say from the set of the speed restrictions which can be displayed by the traffic management system 650.

In accordance with this embodiment, the adaptation rule, known from the Q-learning method, for calculating the optimization function OFQ is extended by a risk control function $\aleph^{\kappa}(.)$, which takes account of the risk.

In turn, the risk control parameter κ is prescribed in accordance with the strategy from the first exemplary embodiment in the interval of $[-1 \leq \kappa \leq 1]$, and represents the risk which a user wishes to run in the application with regard to the control strategy to be determined.

The following evaluation function OFQ is used in accordance with this exemplary embodiment:

$$\mathrm{OFQ} = Q\left(x;\, w^{a}\right), \tag{35}$$

- $x = (v;\, \rho;\, q)$ denoting a state of the traffic system,
- a denoting a speed restriction from the action space A of all speed restrictions which can be displayed by the traffic management system 650, and
- w^a denoting the weights of the neural network which belong to the speed restriction a.

The following adaptation step is executed in Q-learning in order to determine the optimum weights w^a of the neural network:

$$w_{t+1}^{a_t} = w_t^{a_t} + \eta_t \cdot \aleph^{\kappa}\left(d_t\right) \cdot \nabla Q\left(x_t;\, w_t^{a_t}\right) \tag{36}$$

using the abbreviation:

$$d_t = r\left(x_t,\, a_t,\, x_{t+1}\right) + \gamma \max_{a \in A} Q\left(x_{t+1},\, w_t^{a}\right) - Q\left(x_t,\, w_t^{a_t}\right) \tag{37}$$

- x_t, x_{t+1} denoting in each case a state of the traffic system in accordance with rule (34),
- a_t denoting an action, that is to say a speed restriction which can be displayed by the traffic management system 650,
- γ denoting a prescribable reduction factor,
- w_t^{at} denoting the weighting vector belonging to the action a_t, before the adaptation step,

- $w_{t+1}^{a_t}$ denoting the weighting vector belonging to the action a_t, after the adaptation step,
- η_t ($t = 1, \ldots$) denoting a prescribable step size sequence,
- $\kappa \in [-1; 1]$ denoting a risk control parameter,
- $\aleph^\kappa$ denoting a risk control function $\aleph^\kappa(\xi) = (1 - \kappa\mathrm{sign}(\xi))\xi$,
- $\nabla_Q(\ ;\ )$ denoting the derivative of the neural network with respect to its weights, and
- $r(x_t, a_t, x_{t+1})$ denoting a gain upon the transition in state from the state x_t to the subsequent state x_{t+1}.

An action a_t can be selected at random from the possible actions a_t during learning. It is not necessary in this case to select the action a_t which has led to the largest evaluation variable.

The adaptation of the weights has to be performed in such a way that not only is a traffic control achieved which is optimized in terms of the expectation of the optimization function, but that also account is taken of a variance of the control results.

This is particularly advantageous since the state vector x(t) models the actual system of traffic only inadequately in some aspects, and so unexpected disturbances can thereby occur. Thus, the dynamics of the traffic, and therefore of its modeling, depend on further factors such as weather, proportion of trucks on the road, proportion of mobile homes, etc., which are not always integrated in the measured variables of the state vector x(t). In addition, it is not always ensured that the road users immediately implement the new speed instructions in accordance with the traffic management system.

A control phase on the real system in accordance with the traffic management system takes place in accordance with the following steps:

1. The state x_t is measured at the instant t at various points in the traffic system of traffic and yields a state vector $x(t): = (v(t), \rho(t), q(t))$.
2. A value of the optimization function is determined for all possible actions a_t, and that action a_t with the highest evaluation in the optimization function is selected.

Although modifications and changes may be suggested by those skilled in the art to which this invention pertains, it is the intention of the inventors to embody within the patent warranted hereon all changes and modifications that may reasonably and properly come under the scope of their contribution to the art. - -

~~{Description}~~ **[Substitute specification:]**

~~{Method and arrangement}~~[- - METHOD AND ARRANGEMENT FOR DETERMINING A SEQUENCE OF ACTIONS FOR A SYSTEM

BACKGROUND OF THE INVENTION

Field of the Invention:

This invention generally pertains to systems having states, and in particular to methods] for determining a sequence of actions for ~~{a system which has states, }~~**[such systems.**

Discussion of the Related Art:

A generalized method and arrangement for determining a sequence of actions for a system having states, wherein] a transition in state between two states ~~{being}~~ **[is]** performed on the basis of an action**[, is discussed by Neuneier in *Enhancing Q-Learning for Optimal Asset Allocation*, appearing in the Proceedings of the Neural Information Processing Systems, NIPS 1997. Neuneier describes a financial market as an example of]** ~~{The invention relates to a method and an arrangement for determining a sequence of actions for}~~ a system which has states**[. His]**~~{, a transition in state between two states being performed on the basis of an action.~~

~~Such a method and such an arrangement are known from [1].~~

~~A financial market is described in [1] as an example for such a system which has states.~~

~~The}~~ system is described as a Markov **[Decision Problem (MDP).]** ~~{decision problem (MDP). The structure of a system which can be described as a Markov decision problem is illustrated in Figure 2.~~

~~The system 201 is in a state xt at an instant t. The state xt can be observed by an observer of the system. On the basis of an action at from a set in the state xt of possible actions, at (A(xt), the system makes a transition with a certain probability into a subsequent state xt+1 at a subsequent instant t+1.~~

~~This is illustrated diagrammatically in Figure 2 by a loop. An observer 200 perceives 202 observable variables concerning the state xt and takes a decision via an action 203 with which it acts on the system 201. The system 201 is usually subject to the interference 205.~~

~~Furthermore, the observer 200 obtains a gain rt 204~~

~~which is a function of the action at 203 and the original state xt at the instant t as well as of the subsequent state xt+1 of the system at the subsequent instant t+1.~~

~~The gain rt can assume a positive or negative skalar value, depending on whether the decision leads, with regard to a prescribable criterion, to a positive or negative system development, in [1] to an increase in capital stock or to a loss.~~

~~In a further time step, the observer 200 of the system 201 decides on the basis of the observable variables 202, 204 of the subsequent state xt+1 in favor of a new action at+1, etc.~~

~~A sequence of~~

~~State: xt (X~~

~~Action: at (A(xt)~~

~~Subsequent state: xt+1 (X~~

~~Gain rt = r(xt, at, xt+1) ((~~

~~etc. describes a trajectory of the system which is evaluated by a performance criterion which accumulates the individual gains rt over the instants t. It is assumed by way of simplification in a Markov decision problem that the state xt and the action at all contain information for the purpose of describing a transition probability p(xt+1(*) of the system from the state xt to the subsequent state xt+1.~~

~~In formal terms, this means that:~~

~~p(xt+1(xt,at) denotes a transition probability for the subsequent state xt+1 for a given state xt and given action at.~~

~~In a Markov decision problem, future states of the system 201 are thus not a function of states and actions which lie further in the past than one time step.}~~

The characteristics of a Markov ~~{decision problem}~~ **[Decision Problem]** are represented below by way of summary:

X		set of possible states of the system,
		e.g. X = $\Re^{\{()m}$,
$A(x_t)$		set of possible actions in the state
~~{p(xt+1(xt}~~ **[p(x_{t+1} | x_t]**,a_t)	x_t	
$r(x_t, a_t, x_{t+1})$	gain with expectation $R(x_t, a_t)$.	

Starting from observable variables, the variables denoted below as training data, the aim is to determine a strategy, that is to say a sequence of functions

$$\pi = \{\mu_0, \mu_1, K , \mu_T\}, \tag{3}$$

which at each instant t map each state into an action rule, that is to say action

$$\mu_t(x_t) = a_t \tag{4}$$

Such a strategy is evaluated by an optimization function.

The optimization function specifies the expectation, the gains accumulated **[ ]**over time at a given strategy π~~{()~~, and a start state x_0.

The so-called Q-learning method is described ~~{in [1]}~~ **[by Neuneier]** as an example of a method of approximative dynamic programming.

An optimum evaluation function V*(x) is defined by

$$V^{*}(x) = \max_{\pi} V^{\pi}(x) \qquad \forall x \in X \tag{5}$$

where

$$V^{\pi}(x) = E\left[\sum_{t=0}^{\infty} \gamma^{t} r(x_t, \mu_t, x_{t+1}) | x_0 = x\right], \qquad (6)$$

γ {~~(~~}denoting a prescribable reduction factor which is formed in accordance with the following rule:

$$\gamma = \frac{1}{1+z}, \qquad (7)$$

$$z \in \Re^{+}. \qquad (8)$$

A Q-evaluation function $Q^*(x_t,a_t)$ is formed within the Q-learning method for each pair (state x_t, action a_t) in accordance with the following rule:

$$\begin{aligned} Q^*(x_t, a_t) := \quad & \sum_{x \in X} p(x_{t+1}|x_t, a_t) \cdot r_t + \\ & +\gamma \cdot \sum_{x \in X} p(x|x_t, a_t) \cdot \max_{a \in A}\left(Q^*(x, a)\right) \end{aligned}$$

(9)

On the basis respectively of the tupel (x_t, x_{t+1}, a_t, r_t), the Q-values Q^* (x,a) are [ ]adapted in the k+1 th iteration in accordance with the following learning rule with a prescribed learning rate $\eta_{\{\cancel{t}\}k}$ in accordance with the following rule:

$$Q_{k+1}(x_t, a_t) = (1 - \eta_k)Q_k(x_t, a_t) + \eta_k\left(r_t + \gamma \max_{a \in A}\left(Q_k(x_{t+1}, a)\right)\right). \quad (10)$$

Usually, the so-called Q-values $Q^*(x,a)$ are approximated for various actions {~~a~~} by a function approximator in each case, for example a neural network or {~~else~~} a polynomial classifier, with a weighting vector w^a, which contains weights of the function approximator.

A function approximator is {~~to be understood as~~}, for example, a neural network, a polynomial classifier or {~~else~~} a combination of a neural network with a polynomial classifier.

It therefore holds that:

$$Q^*(x, a) \approx Q\left(x; w^a\right). \tag{11}$$

Changes in the weights in the weighting vector w^a are based on a temporal difference d_t which is formed in accordance with the following rule:

$$d_t := r\left(x_t, a_t, x_{t+1}\right) + \gamma \max_{a \in A} Q\left(x_{t+1}; w_k^a\right) - Q\left(x_t; w_k^{a_t}\right) \tag{12}$$

The following adaptation rule for the weights of the neural network, which are included in the weighting vector w^a, follows for the Q-learning method with the use of a neural network:

$$w_{k+1}^{a_t} = w_k^{a_t} + \eta_k \cdot d_t \cdot \nabla Q\left(x_t; w_k^{a_t}\right). \tag{13}$$

The neural network representing the system of a financial market~~{,}~~ as described ~~{in [1],}~~ **[by Neuneier]** is trained using the training data which describe information on changes in prices on a financial market as time series values.

A further method of approximative dynamic programming~~{, the so-called TD(()-learning method, is known from [2] and is explained in more detail in conjunction with an exemplary embodiment.}~~ **[is the so-called TD(λ) learning method. This method is discussed in R.S. Sutton*s, *Learning To Predict By The Method Of Temporal Differences*, appearing in Machine Learning, Chapter 3, pages 9 - 44, 1988.]**

Furthermore, it is known from ~~{[3] which}~~ **[M. Heger*s, *Risk and Reinforcement Learning: Concepts and Dynamic Programming*, ZKW Bericht No. 8/94, Zentrum f*r Kognitionswissenschaften [Center for Cognitive Sciences], Bremen University, December 1994, that]** risk is associated with a strategy π ~~{(}~~and an initial **[**

]state x_t. A method for risk ~~{avoidment is likewise known from [3]}~~ **[avoidance is also discussed by Hager, cited above]**.

The following optimization function, which is also referred to as an expanded Q-function ~~{Q((xt}~~ **[**$\underline{Q}^{\pi}$**(x$_t$]**, a_t), is used in the **[Hager]** method ~~{known from [3]}~~:

maximize

$$\left(\underline{Q}^{\pi}(x_t, a_t) := r(x_t, a_t, x_{t+1}) + \inf_{\substack{x_0, x_1, K \\ p(x_0, x_1, K) > 0}} \left\{ \sum_{k=1}^{\infty} \gamma^k r(x_k, \pi(x_k), x_{k+1}) \right\}\right) \tag{14}$$

The expanded Q-function ~~{Q((xt}~~ **[**$\underline{Q}^{\pi}$**(x$_t$]**, a_t) describes the worst case if the action a_t is executed in the state x_t and the strategy π ~~{(}~~is followed thereupon.

The optimization function ~~{Q((xt}~~ **[**$\underline{Q}^{\pi}$**(x$_t$]**, a_t) for

$$\underline{Q}^{*}(x_t, a_t) := \max_{\pi \in \Pi} \underline{Q}^{\pi}(x_t, a_t) \tag{15}$$

~~{,}~~

is given by the following rule:

$$\underline{Q}^{*}(x_t, a_t) = \min_{\substack{x \in X \\ p(x_{t+1} | x_t, a_t) > 0}} \left(r(x_t, a_t, x) + \gamma \cdot \max_{a \in A} \underline{Q}^{*}(x, a) \right). \tag{16}$$

A substantial disadvantage of this mode of procedure is ~~{to be seen in}~~ that only the worst case is taken into account when finding the strategy. However, this **[inadequately]** reflects the requirements of the most varied technical systems ~~{only to an inadequate extent.}~~**[.]**

~~{Furthermore, it is known from [4] to formulate}~~ **[In *Dynamic Programming and Optimal Control*, Athena Scientific, Belmont, MA, 1995, D.P. Bertsekas formulates]** access control for a communications network **[** **]**and ~~{the}~~ routing within the communications network as a problem of dynamic **[** **]**programming.

~~{The}~~ **[Therefore, the present]** invention is ~~{therefore}~~ based on the problem of specifying a method and ~~{an arrangement}~~ **[system]** for determining a sequence of actions ~~{for a system,}~~ in which **[the]** method or ~~{action}~~ **[sequences of actions achieve]** an increased flexibility in determining the strategy ~~{is achieved.}~~ **[needed.]**

~~{The problem is solved by the method and by the arrangement in accordance with the features of the independent patent claims.~~

~~In a method for computer-aided determination of a sequence of actions for a system which has states, a transition in state between two states being performed on the basis of an action, the determination of the sequence of actions is performed in such a way that a sequence of states resulting from the sequence of actions is optimized with regard to a prescribed optimization function, the optimization function including a variable parameter with the aid of which it is possible to set a risk which the resulting sequence of states has with respect to a prescribed state of the system.~~

~~An arrangement for determining}~~ **[In a method for computer-aided determination of]** a sequence of actions for a system which has states, a transition in state between two states being performed on the basis of an action, ~~{has a processor which is set up in such a way that}~~ the determination of the sequence of actions ~~{can be}~~ **[is]** performed in such a way that a sequence of states resulting from the sequence of actions is optimized with regard to a prescribed optimization function, the optimization function including a variable parameter with the aid of which it is possible to set a risk which the resulting sequence of states has with respect to a prescribed state of the system.

[A system for determining a sequence of actions for a system which has states, a transition in state between two states being performed on the basis of an action, has a processor which is set up in such a way that the determination of the sequence of actions can be performed in such a way that a sequence of states resulting from the sequence of actions is optimized with regard to a prescribed optimization function, the optimization function including a variable parameter with the aid of which it is possible to set a risk which the resulting sequence of states has with respect to a prescribed state of the system.

Thus, the present invention offers] ~~{It becomes possible for the first time owing to the invention to specify}~~ a method for determining a sequence of actions at a freely prescribable level of accuracy when finding a strategy for a possible closed-loop control or open-loop control of the system, in general for influencing it. **[Hence, the embodiments]** ~~{Preferred developments of the invention follow from the dependent claims.~~
~~The developments}~~ described below are valid both for the method and for the ~~{arrangement, the processor being respectively set up in the development of arrangement in such a way that the development can be implemented.}~~ **[system.]**

~~{In a preferred refinement, a method of approximative}~~ **[Approximative]** dynamic programming is used for the purpose of ~~{determination}~~ **[determina-tion]**, for example a method based on Q-learning or ~~{else}~~ a method based on TD(λ~~{()}~~**[)]**-learning.

Within Q-learning, the optimization function OFQ is preferably formed in accordance with the following rule:

$$OFQ = Q\left(x;\ w^{a}\right),$$

- ~~{()}~~x denoting a state in a state space X
- ~~{()}~~a denoting an action from an action space A, and
- ~~{()}~~w^a denoting the weights of a function approximator which belong to the action a.

The following adaptation step is executed during Q-learning in order to determine the optimum weights w^a of the function approximator:

$$w_{t+1}^{a_t} = w_t^{a_t} + \eta_t \cdot \aleph^{\kappa}\left(d_t\right) \cdot \nabla Q\left(x_t; w_t^{a_t}\right)$$

with the abbreviation

$$d_t = r\left(x_t, a_t, x_{t+1}\right) + \gamma \max_{a \in A} Q\left(x_{t+1}, w_t^a\right) - Q\left(x_t, w_t^{a_t}\right)$$

- ~~{()~~x_t, x_t+1 respectively denoting a state in the state space X,
- ~~{()~~a_t denoting an action from an action space A,
- γ ~~{(()~~denoting a prescribable reduction factor,
- $w_t^{a_t}$ [redacted] ~~{()~~denoting the weighting vector associated with the action a_t before the adaptation step,
- $w_{t+1}^{a_t}$ [redacted] ~~{()~~denoting the weighing vector associated with the action a_t after the adaptation step,
- η~~{(()~~$_t$ (t = 1, ...) denoting a prescribable step size sequence,
- $\kappa \in$ ~~{((()~~[-1; 1] denoting a risk monitoring parameter,
- $\aleph^\kappa$ ~~{((()~~denoting a risk monitoring function $\aleph^\kappa$ **(ξ[) = (1*- κsign(ξ))ξ,]**~~{((() = (1 - (sign(()}(,~~

~~(()~~

- **∇Q(*;*)** denoting the derivation of the function approximator according to its weights, and
- ~~{()~~$r(x_t, a_t, x_{t+1})$ denoting a gain upon the transition of state from the state x_t to the subsequent state x_{t+1}.

The optimization function is preferably formed in accordance with the following **[**

]rule within the TD(λ~~{()}~~**[)]**-learning method:

OFTD = J(x;w)

- ~~{()~~x denoting a state in a state space X,
- ~~{()~~a denoting an action from an action space A, and
- ~~{()~~w denoting the weights of a function approximator.

The following adaptation step is executed during TD(λ~~{()}~~**[)]**-learning in order to determine the optimum weights w of the function approximator:

$w_{t+1} = w_t + \eta_{\cancel{\{((t * (()}[t} \mathbf{*]}\ \aleph^{\kappa}(d_t) * z_t$

with the abbreviations

$d_t = r(w_t, a_t, x_{t+1}) + \gamma\cancel{\{()}J(x_{t+1}; w_t) - J(x_t; w_t),$

$z_t = \lambda\ \cancel{\{(* ()}\mathbf{[*]}\ \gamma * z_{t-1} + \nabla\cancel{\{()}J(x_t; w_t),$

$z_{-1} = 0$

- ~~{()~~x_t, x_{t+1} respectively denoting a state in the state space X,
- ~~{()~~a_t denoting an action from an action space A,
- γ ~~{(()~~denoting a prescribable reduction factor,
- ~~{()~~w_t denoting the weighting vector before the adaptation step,
- ~~{()~~w_{t+1} denoting the weighting vector after the adaptation step,
- $\eta_{\cancel{\{(()}t}$ (t = 1, ...) denoting a prescribable step size sequence,
- κ ∈ ~~{((()~~[-1; 1] denoting a risk monitoring parameter,
- $\aleph^{\kappa}$ ~~{((()~~denoting a risk monitoring function $\aleph^{\kappa}$ **(ξ[) = (1*- κsign(ξ))ξ,]**~~{((() = (1 - (sign(()(,~~

~~(()~~

- **∇J(*;*)** denoting the derivation of the function approximator according to its **[**

  ]weights, and
- ~~{()~~$r(x_t, a_t, x_{t+1})$ denoting a gain upon the transition of state from the state x_t to the subsequent state x_{t+1}.

[<u>SUMMARY OF THE INVENTION</u>

It is an object of the present invention to provide a technical system and method for determining a] ~~{The system is preferably a technical system of which before the determination measured values are measured which are used in determining the}~~ sequence of actions **[using measured values.**

It is another object of the present invention to provide a technical system and method that]~~{.~~
~~The technical system }~~ can be subjected to open-loop control or ~~{else}~~ closed-loop control with the use of ~~{the}~~ **[a]** determined sequence of actions.

~~{The system is preferably}~~ **[It is a further object of the invention to provide a technical system and method]** modeled as a Markov ~~{decision problem.}~~ **[Decision Problem.]**

~~{The method or the arrangement is preferably}~~ **[It is an additional object of the invention to provide a technical system and method that can be]** used in a traffic management system ~~{or}~~**[.**

It is yet another object of the invention to provide a technical system and method that can be used] in a communications system, ~~{the}~~ **[such that a]** sequence of actions ~~{being used in a communications network}~~ **[is used]**to carry out access control ~~{or a routing, that is to say a path allocation.}~~**[, routing or path allocation.]**

~~{Furthermore, the system can be a financial market which is modeled by a Markov decision problem, the change in the financial market, for example the}~~ **[It is yet a further object of the invention to provide a technical system and**
method for a financial market modeled by a Markov Decision Problem, wherein a] change in an
{
}index of stocks**[,]** or ~~{else}~~ **[a change in]** a rate of exchange on a foreign exchange market ~~{being analyzed by using the method and/or the arrangement and it being}~~**[, makes it]**possible to intervene in the market in accordance with ~~{the}~~ **[a]** sequence of determined actions.

[These and other objects of the invention will be apparent from a careful review of the following detailed description of the preferred embodiments, which is to read in conjunction with a review of the accompanying drawing figures.

BRIEF DESCRIPTION OF THE DRAWINGS] ~~{Exemplary embodiments of the invention are illustrated in the figures and explained in more detail below.}~~

Figure 1 shows a flowchart ~~{in which individual}~~ **[of]** method steps ~~{of the first exemplary embodiment are illustrated}~~ **[according to the present invention]**;

Figure 2 shows a ~~{sketch of a}~~ system ~~{which can be}~~ modeled as a Markov ~~{decision problem}~~ **[Decision Problem]**;

Figure 3 shows a ~~{sketch of a}~~ communications network ~~{in which}~~ **[wherein]** access control is carried out in a switching unit **[according to the present invention]**;

Figure 4 shows a ~~{symbolic sketch of a}~~ function approximator ~~{with the aid of which a method of}~~ **[for]** approximative dynamic programming ~~{is implemented}~~ **[according to the present invention]**;

Figure 5 shows ~~{a further sketch of}~~ a plurality of function approximators ~~{with the aid of which}~~ **[for]** approximative dynamic programming ~~{is implemented}~~ **[according to the present invention]**; and

Figure 6 shows a ~~{sketch of a}~~ traffic management system ~~{which is}~~ subjected to closed-loop control in accordance with ~~{an exemplary embodiment.}~~ **[the present invention.]**

~~{First exemplary embodiment: access control and routing.}~~ [DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS]

~~{Figure 3 shows }~~**[Figure 1 shows a flowchart according to the present invention, in which individual method steps of a first embodiment are provided, which will be discussed later.**

Figure 2 shows the structure of a typical Markov Decision Problem method.

The system 201 is in a state x_t at an instant t. The state x_t can be observed by an observer of the system. On the basis of an action a_t from a set in the state x_t of possible actions, $a_t \in A(x_t)$, the system makes a transition with a certain probability into a subsequent state x_t+1 at a subsequent instant t+1.

As illustrated diagrammatically in Figure 2 by a loop, an observer 200 perceives 202 observable variables concerning the state x_t and takes a decision via an action 203 with which it acts on the system 201. The system 201 is usually subject to the interference 205.

The observer 200 obtains a gain r_t 204

$$r_t = r(x_t, a_t, x_{t+1}) \in \Re, \tag{1}$$

which is a function of the action a_t 203 and the original state x_t at the instant t as well as of the subsequent state x_t+1 of the system at the subsequent instant t+1.

The gain r_t can assume a positive or negative scalar value depending on whether the decision leads, with regard to a prescribable criterion, to a positive or negative system development, to an increase in capital stock or to a loss.

In a further time step, the observer 200 of the system 201 decides on the basis of the observable variables 202, 204 of the subsequent state x_{t+1} in favor of a new action a_{t+1}, etc.

A sequence of

State:	x_t	$\in$	X
Action:	a_t	$\in$	$A(x_t)$
Subsequent state:	x_t+1	$\in$	X
Gain	$r_t = r(x_t, a_t, x_{t+1})$	$\in$	$\Re$

describes a trajectory of the system which is evaluated by a performance criterion which accumulates the individual gains r_t over the instants t. It is assumed by way of simplification in a Markov Decision Problem that the state x_t and the action a_t all contain information for the purpose of describing a transition probability $p(x_{t+1}|*)$ of the system from the state x_t to the subsequent state x_{t+1}.

In formal terms, this means that:

$$p\left(x_{t+1} \middle| x_t, K\ , x_0, a_t, K\ , a_0\right) = p\left(x_{t+1} \middle| x_t, a_t\right). \qquad (2)$$

$p(x_{t+1}|x_t,a_t)$ denotes a transition probability for the subsequent state x_{t+1} for a given state x_t and given action a_t.

In a Markov Decision Problem, future states of the system 201 are thus not a function of states and actions which lie further in the past than one time step.

Figure 3 shows an embodiment of the present invention involving an access control and routing system, such as] a communications network 300{~~, which~~}**[.**

The communications network 300] has a multiplicity of switching units 301a, 301b, ..., 301i, ... 301n, which are interconnected via connections 302a, 302b, 302j, ... 302m. **[A]** {~~Furthermore, a~~} first terminal 303 is connected to a first switching unit 301a. From the first terminal 303, the first switching unit 301a is sent a request message 304 which requests preservation of a prescribed bandwidth within the communications network 300 for the purpose of transmitting data{~~(~~}**[, such as]** video data{~~,~~} **[or]** text data{~~)~~}.

It is determined in the first switching unit 301a in accordance with a strategy described below**[,]** whether the requested bandwidth is available in the {~~communications~~} **[communi-cations]** network 300 on a specified, requested connection {~~(step 305)~~} **[instep 305]**. {
}The request is refused {~~(step 306)~~} **[instep 306]** if this is not the case. {

}If sufficient bandwidth is available, it is checked in {~~a further~~} checking step {~~(step 307)~~} **[307]** whether the bandwidth can be reserved.

The request is refused {~~(step 308)~~} **[in step 308]** if this is not the case. { }Otherwise, the first switching unit 301a selects a route from the first switching unit 301a via further switching units 301i to a second terminal 309 with which the first terminal 303 wishes to communicate, and a connection is initialized {~~(step 310)~~} **[in step 310]**.

The starting point below is a communications network 300 which comprises a set of switching units

$$N = \{1, K , n, K , N\} \tag{17}$$

and a set of physical connections

$$L = \{1, K , 1, K , L\} , \tag{18}$$

a physical connection l having a capacity of B(l) bandwidth units.

A set

$$M = \{1, K , m, K , M\} \tag{19}$$

of different types of service m are available, a type of service m being characterized by

- {~~()~~}a bandwidth requirement b(m),
- {~~()~~}an average connection time $\frac{1}{V(m)}$, [redacted] and
- {~~()~~}a gain c(m) which is obtained whenever a call request of the corresponding type of service m is accepted.

The gain c(m) is given by the amount of money which a network operator of the communications network 300 bills a subscriber for a connection of the type of service. Clearly, the gain c(m) reflects different priorities, which can be prescribed by the network operator and which he associates with different services.

A physical connection 1 can simultaneously provide any desired combination of communications connections as long as the bandwidth used for the communications connections does not exceed the bandwidth available overall for the physical connection.

If a new communications connection of type m is requested between a first node i and a second node j (terminals are also denoted as nodes), the requested communications connection can, as represented above, either be accepted or be refused. {
}If the communications connection is accepted, a route is selected from a set of prescribed routes. This selection is denoted as a routing. b(m) bandwidth units are used in the communications connection of type m for each physical connection along **[**
]the selected route for the duration of the connection.

Thus, during access control{~~(~~}**[, also referred to as]** call admission control{~~)~~}, a route can be selected within the communications network 300 only when the selected route has sufficient bandwidth available. {
}The aim of the access control and of the routing is to maximize a long term gain which is obtained by acceptance of the requested connections.

At an instant t, the technical system which is the communications network 300 is in a state x_t which is described by a list of routes via existing connections, by means of which lists it is shown how many connections of which type of service are using the respective routes at the instant t.

Events w, by means of which a state x_t could be transferred into a subsequent state x_{t+1}, are the arrival of new connection request messages, or else the {~~termination~~} **[termina-tion]** of a connection existing in the communications network 300.

In this {~~exemplary~~} embodiment, an action a_t at an instant t**[,]** owing to a connection request is the {

}decision as to whether a connection request is to be accepted or refused and, if the connection is accepted, the selection of the route through the communications network 300.

The aim is to determine a sequence of actions, that is to say clearly to determine the learning of a strategy with actions relating to a state x_t in such a way that the following rule is maximized:

$$E\left(\sum_{k=0}^{\infty} e^{-\beta t_k} \cdot g\left(x_{t_k}, \omega_k, a_{t_k}\right)\right), \tag{20}$$

- {()E{.} denoting an expectation,
- {()t_k denoting an instant at which a kth event takes place,
- $g\left(x_{t_k}, \omega_k, a_{t_k}\right)$ [redacted] {(). denoting the gain which is associated with the kth event, and
- β {{-()denoting a reduction factor which evaluates an immediate gain as being more valuable than a gain at instants lying further in the future.

Different implementations of a strategy lead normally to different overall gains G:

$$G = \sum_{k=0}^{\infty} e^{-\beta t_k} \cdot g\left(x_{t_k}, \omega_k, a_{t_k}\right). \tag{21}$$

The aim is to maximize the expectation of the overall gain G in accordance with the following rule J:

$$J = E\left\{\sum_{k=0}^{\infty} e^{-\beta t_k} \cdot g\left(x_{t_k}, \omega_k, a_{t_k}\right)\right\}, \tag{22}$$

it being possible to set a risk which reduces the overall gain G of a specific implementation of access control and of a routing strategy to below the expectation.

The TD(λ{()}[)]-learning method is used to carry out the access control and the [ ]routing.{

}

The following target function is used in this ~~{exemplary}~~ embodiment:

$$J^{*}(x_t) = E_{\tau}\left\{e^{-\beta\tau}\right\}E_{\omega}\left\{\max_{a \in A}\left[g(x_t, \omega_t, a) + J^{*}(x_{t+1})\right]\right\}, \tag{23}$$

- ~~{()}~~A denoting an action space with a prescribed number of actions which are respectively available in a state x_t,
- τ ~~{(-()}~~denoting a first instant at which a first event ω ~~{()}~~occurs, and
- ~~{()}~~x_{t+1} denoting a subsequent state of the system.

An approximated value of the target value $J^*(x_t)$ is learned and stored by employing a function approximator 400 (compare Figure 4) with the use of training data.

Training data are data previously measured in the communications network 300 and relating to the behavior of the communications network 300 in the case of incoming connection requests 304 and of termination of messages. This time sequence of states is stored, and these training data are used to train the function approximator 400 in accordance with the learning method described below.

A number of connections of in each case one type of service m on a route of the communications network 300 serve in each case as input variable of the function approximator 400 for each input 401, 402, 403 of the function approximator 400. **[**

]These are represented ~~{symbolically}~~ in Figure 4 by blocks 404, 405, 406. {

}An approximated target value $\tilde{J}$ [redacted] of the target value J^* is the output variable of the function approximator 400.

Figure 5 shows a detailed representation of ~~{the}~~ **[a]** function approximator 500, which ~~{in this case}~~ has several component function approximators 510, 520 ~~{of the function approximator 500}~~. **[**

]One output variable is the approximated target value $\tilde{J}$ [redacted], which is formed in accordance with the following rule:

$$\tilde{J}(x_t, \Theta) = \sum_{l=1}^{L} \tilde{J}^{(l)}\left(x_t^{(l)}, \Theta_t^{(l)}\right). \tag{24}$$

The input variables of the component function approximators 510, 520, which are present at the inputs 511, 512, 513 of the first component function approximator 510, or at the inputs 521, 522 and 523 of the second component function approximator 520 are, in turn, respectively a number of types of service of a type m in a physical connection r in each case, symbolized by blocks 514, 515, 516 for the first component function approximator, and 524, 525 and 526 for the second component function approximator 520.

Component output variables 530, 531, 532, 533 are fed to an adder unit 540, and the approximated target variable $\tilde{J}$ [redacted] is formed as output variable of the adder unit.

Let it be assumed that the communications network 300 is in the state x_{t_k} [redacted] and that a request message with which a type of service m of class m is requested for a connection { }between two nodes i, j reaches the first switching unit 301a.

A list of permitted routes between the nodes i and j is denoted by R(i, j), and a list of all possible routes is denoted by

$$\tilde{R}\left(i, j, x_{t_k}\right) \subset R(i, j) \tag{25}$$

as a subset of the routes R(i, j) which could implement a possible connection with regard to the available and requested bandwidth.

For each possible route r, $r \in \tilde{R}\left(i, j, x_{t_k}\right)$ [redacted], a subsequent state $x_{t_k} + 1\left(x_{t_k}, \omega_k, r\right)$ [redacted] is determined which results from the fact that the connection request 304 is accepted and the connection on the route r is made available to the requesting first terminal 303.

This is illustrated in Figure 1 as ~~{second}~~ step ~~{(step 102)}~~ **[102]**, the state of the system and the respective event being respectively determined in ~~{a first}~~ step ~~{(step 101)}~~ **[101]**. {

}A route r* to be selected is determined in ~~{a third}~~ step ~~{(step 103)}~~ **[103]** in accordance with the following rule:

$$r^* = \arg \max_{r \in \tilde{R}\left(i, j, x_{t_k}\right)} \tilde{J}\left(x_{t_k+1}\left(x_{t_k}, \omega_k, r\right), \Theta_t\right). \tag{26}$$

A check is made in ~~{a further}~~ step ~~{(step 104)}~~ **[104]** as to whether the following rule is fulfilled:

$$c(m) + \tilde{J}\left(x_{t_k+1}\left(x_{t_k}, \omega_k, r^*\right), \Theta_t\right) < \tilde{J}\left(x_{t_k}, \Theta_t\right). \tag{27}$$

If this is the case, the connection request 304 is rejected ~~{(step 105)}~~ **[in step 105]**, otherwise the connection is accepted and *switched through* to the node j along the selected route r* ~~{(step 106)}~~ **[in step 106]**.

Weights of the function approximator 400, 500 which are adapted in the TD(λ~~{()}~~**[)]**-learning method to the training data, are stored in a parameter vector θ ~~{(}~~for an instant t in each case, such that an optimized access control and an optimized routing are achieved.

During the training phase, the weighting parameters are adapted to the training data applied to the function approximator.

A risk parameter κ ~~{(}~~is defined with the aid of which a desired risk, which the system has with regard to a prescribed state owing to a sequence of actions and states, can be set in accordance with the following rules:

-1 ≤ κ **~~{(-(}~~< 0:**	risky learning,
κ ~~{(}~~= 0:	neutral learning with regard to the risk,
0 < κ **~~{(}~~< 1:**	risk-avoiding learning,
κ ~~{(}~~= 1:	worst-case learning.

Furthermore, a prescribable parameter $0 \leq \lambda \leq$ ~~{((()}~~1 and a step size sequence γ~~{()}~~k are prescribed in the learning method.

The weighting values of the weighting vector Θ ~~{()}~~are adapted to the training data on the basis of each event ω_{t_k} [redacted] in accordance with the following adaptation rule:

$$\Theta_k = \Theta_{k-1} + \gamma_k \aleph^{\kappa}(d_k) z_t , \tag{28}$$

in which case

$$d_k = e^{-\beta(t_k - t_{k-1})}\left(g\left(x_{t_k}, \omega_k, a_{t_k}\right) + \tilde{J}\left(x_{t_k}, \Theta_{k-1}\right)\right) - \tilde{J}\left(x_{t_{k-1}}, \Theta_{k-1}\right) \tag{29}$$

$$z_t = \lambda e^{-\beta(t_{k-1} - t_{k-2})} z_{t-1} + \nabla_{\Theta} \tilde{J}\left(x_{t_{k-1}}, \Theta_{k-1}\right), \tag{30}$$

and

$$\aleph^{\kappa}(\xi) = \left(1 - \kappa \mathrm{sign}(\xi)\right)\xi . \tag{31}$$

It is assumed that: $z_{-1} = 0$.

The function

$$g\left(x_{t_k}, \omega_k, a_{t_k}\right) \tag{32}$$

denotes the immediate gain in accordance with the following rule:

]

Thus, as described above, a sequence of actions is determined with regard to a connection request such that a connection request is either rejected or accepted on the basis of an action. The determination is performed taking account of an optimization function in which the risk can be set by means of a risk control parameter κ ∈ ~~{()}~~[-1; 1] in a variable fashion.

~~{Second exemplary embodiment: Traffic management system}~~ **[Figure 6 shows an embodiment of the present invention in relation to a traffic management system]**

~~{Figure 6 shows a }~~**[A]** road 600 on which automobiles 601, 602, 603, 604, 605 and 606 are being **[**

]driven.{

} Conductor loops 610, 611 integrated into the road 600 receive electric signals in a known way and feed the electric signals 615, 616 to a computer 620 via an input/output interface 621. In an analog-to-digital converter 622 connected to the input/output interface 621, the electric signals are digitized into a time series and stored in a memory 623, which is connected by a bus 624 to the analog-to-digital converter 622 and a processor 625. Via the input/output interface 621, a traffic management system 650 is fed control signals 651 from which it is possible to set a prescribed speed stipulation 652 in the traffic management system 650, or else further particulars of traffic regulations, which are displayed via the traffic management system 650 to drivers of the vehicles 601, 602, 603, 604, 605 and 606.

The following local state variables are used in this case for the purpose of traffic modeling:

- ~~{(}~~traffic flow rate v,
- ~~{(}~~vehicle density ρ (ρ ~~{((}~~= number of vehicles per kilometer $\frac{\mathrm{Fz}}{\mathrm{km}}$ [redacted]).
- ~~{(}~~traffic flow q (q = number of vehicles per hour $\frac{\mathrm{Fz}}{\mathrm{h}}$ [redacted], (q= v * ρ~~{(}~~)), and
- ~~{(}~~speed restrictions 652 displayed by the traffic management system 650 at an instant in each case.

The local state variables are measured as described above by using the conductor loops 610, 611.

These variables (v(t), ρ~~{(}~~(t), q(t)) therefore represent a state of the technical system of *traffic* at a specific instant t.

In this {~~exemplary~~} embodiment, the system is therefore a traffic system which is controlled by using the traffic management system 650**[, and]**{~~.~~

~~In this second exemplary embodiment,~~ } an extended Q-learning method is described as method of approximative dynamic programming.

The state x_t is described by a state vector

$$x(t) = (v(t), \rho(t), q(t)) \,. \qquad (34)$$

The action a_t denotes the speed restriction 652, which is displayed at the instant t by the traffic management system 650. {

}The gain $r(x_t, a_t, x_{t+1})$ describes the quality of the traffic flow which was measured between the instants t and t+1 by the conductor loops 610 and 611. **[**

]In this {~~second exemplary~~} embodiment, $r(x_t, a_t, x_{t+1})$ denotes

- {~~(~~}the average speed of the vehicles in the time interval [t, t + 1]

or

- {~~(~~}the number of vehicles which have passed the conductor loops 610 and 611 in the time interval [t, t + 1]

or

- {~~(~~}the variance of the vehicle speeds in the time interval [t, t + 1],

or

- {~~(~~}a weighted sum from the above variables.

A value of the optimization function OFQ is determined for each possible action a_t, that is to say for each speed restriction which can be displayed by the traffic management system 650, an estimated value of the optimization function OFQ being realized in each case as a neural network.

This results in a set of evaluation variables for the various actions a_t in the **[**

]system state x_t.{

} Those actions a_t for which the maximum evaluation variable OFQ has been determined in the current system state x_t are selected in a control phase from the possible actions a_t, that is to say from the set of the speed restrictions which can be displayed by the traffic management system 650.

In accordance with this {~~exemplary~~} embodiment, the adaptation rule, known from the Q-learning method, for calculating the optimization function OFQ is extended by a risk control function $\aleph^\kappa${~~(()~~}(.), which takes account of the risk.

In turn, the risk control parameter κ {~~()~~}is prescribed in accordance with the strategy from the first exemplary embodiment in the interval of [-1{~~(-(-()~~}[*≤*κ*≤*]1], and represents the risk which a user wishes to run in the application with regard to the control strategy to be determined.

The following evaluation function OFQ is used in accordance with this exemplary embodiment:

$$OFQ = Q\left(x;\ w^a\right), \tag{35}$$

- {~~()~~}x = (v; ρ{~~()~~}; q) denoting a state of the traffic system,
- {~~()~~}a denoting a speed restriction from the action space A of all speed restrictions which can be displayed by the traffic management system 650, and
- {~~()~~}w^a denoting the weights of the neural network which belong to the speed restriction a.

The following adaptation step is executed in Q-learning in order to determine **[**

]the optimum weights w^a of the neural network:

$$w_{t+1}^{a_t} = w_t^{a_t} + \eta_t \cdot \aleph^\kappa\left(d_t\right) \cdot \nabla Q\left(x_t;\ w_t^{a_t}\right) \tag{36}$$

using the abbreviation**[:]**

$$d_t = r\left(x_t, a_t, x_{t+1}\right) + \gamma \max_{a \in A} Q\left(x_{t+1}, w_t^a\right) - Q\left(x_t, w_t^{a_t}\right) \tag{37}$$

- {~~()~~}x_t, x_{t+1} denoting in each case a state of the traffic system in accordance with rule (34),

- ~~{()~~a_t denoting an action, that is to say a speed restriction which can be displayed by the traffic management system 650,
- γ ~~{(()~~denoting a prescribable reduction factor,
- $w_t^{a_t}$ [redacted] ~~{()~~denoting the weighting vector belonging to the action a_t, before the adaptation step,
- $w_{t+1}^{a_t}$ [redacted] ~~{()~~denoting the weighting vector belonging to the action a_t, after the adaptation step,
- η~~{(()~~$_t$ (t = 1, ...) denoting a prescribable step size sequence,
- $\kappa \in$ ~~{((()~~[-1; 1] denoting a risk control parameter,
- $\aleph^\kappa$ ~~{((()~~denoting a risk control function $\aleph^\kappa$ **(ξ[) = (1*-*κsign(ξ))ξ,]**~~{((() = (1 - (sign(())(,~~

~~(()~~

- ∇_Q**(*;*)** denoting the derivative of the neural network with respect to its weights, and
- ~~{()~~$r(x_t, a_t, x_{t+1})$ denoting a gain upon the transition in state from the state x_t to the subsequent state x_{t+1}.

An action a_t can be selected at random from the possible actions a_t during **[**
]learning. It is not necessary in this case to select the action a_t which has led to the largest evaluation variable.

The adaptation of the weights has to be performed in such a way that not only is **[**
]a traffic control achieved which is optimized in terms of the expectation of the optimization function, but that also account is taken of a variance of the control results.

This is particularly advantageous since the state vector x(t) models the actual system of traffic only inadequately in some aspects, and so unexpected disturbances can thereby occur. Thus, the dynamics of the traffic, and therefore of its modeling, depend on further factors such as weather, proportion of trucks on the road, proportion of mobile homes, etc., which are not always integrated in the measured variables of the state vector x(t). In addition, it is not always ensured that the road users immediately implement the new speed instructions in accordance with the traffic management system.

A control phase on the real system in accordance with the traffic management system takes place in accordance with the following steps:

1. The state x_t is measured at the instant t at various points in the traffic system of traffic and yields a state vector $x(t): = (v(t), \rho$~~{()}~~$(t), q(t))$.

2. A value of the optimization function is determined for all possible actions a_t, and that action a_t with the highest evaluation in the optimization function is selected.

~~Abstract~~

~~Method and arrangement for determining a sequence of actions for a system which has states, a transition in state between two states being performed on the basis of an action~~

~~The determination of a sequence of actions is performed in such a way that a sequence of states resulting from the sequence of actions is optmized with regard to a prescribed optimization function. The optimization function includes a variable parameter with the aid of which it is possible to set a risk which the resulting sequence of states has with respect to a prescribed state of the system.}~~ **[Although modifications and changes may be suggested by those skilled in the art to which this invention pertains, it is the intention of the inventors to embody within the patent warranted hereon all changes and modifications that may reasonably and properly come under the scope of their contribution to the art. - -]**

4/PRTS

Description

Method and arrangement for determining a sequence of actions for a system which has states, a transition in state between two states being performed on the basis of an action

The invention relates to a method and an arrangement for determining a sequence of actions for a system which has states, a transition in state between two states being performed on the basis of an action.

Such a method and such an arrangement are known from [1].

A financial market is described in [1] as an example for such a system which has states.

The system is described as a Markov decision problem (MDP). The structure of a system which can be described as a Markov decision problem is illustrated in **Figure 2**.

The system 201 is in a state x_t at an instant t. The state x_t can be observed by an observer of the system. On the basis of an action a_t from a set in the state x_t of possible actions, $a_t \in A(x_t)$, the system makes a transition with a certain probability into a subsequent state x_t+1 at a subsequent instant t+1.

This is illustrated diagrammatically in **Figure 2** by a loop. An observer 200 perceives 202 observable variables concerning the state x_t and takes a decision via an action 203 with which it acts on the system 201. The system 201 is usually subject to the interference 205.

Furthermore, the observer 200 obtains a gain r_t 204

$$r_t = r(x_t, a_t, x_{t+1}) \in \Re, \tag{1}$$

which is a function of the action a_t 203 and the original state x_t at the instant t as well as of the subsequent state x_t+1 of the system at the subsequent instant t+1.

The gain r_t can assume a positive or negative skalar value, depending on whether the decision leads, with regard to a prescribable criterion, to a positive or negative system development, in [1] to an increase in capital stock or to a loss.

In a further time step, the observer 200 of the system 201 decides on the basis of the observable variables 202, 204 of the subsequent state x_{t+1} in favor of a new action a_{t+1}, etc.

A sequence of

State:	x_t	$\in$	X
Action:	a_t	$\in$	$A(x_t)$
Subsequent state:	x_t+1	$\in$	X
Gain	$r_t = r(x_t, a_t, x_{t+1})$	$\in$	$\Re$

etc. describes a trajectory of the system which is evaluated by a performance criterion which accumulates the individual gains r_t over the instants t. It is assumed by way of simplification in a Markov decision problem that the state x_t and the action a_t all contain information for the purpose of describing a transition probability $p(x_{t+1}|\cdot)$ of the system from the state x_t to the subsequent state x_{t+1}.

In formal terms, this means that:

$$p(x_{t+1}|x_t, K, x_0, a_t, K, a_0) = p(x_{t+1}|x_t, a_t). \tag{2}$$

$p(x_{t+1}|x_t, a_t)$ denotes a transition probability for the subsequent state x_{t+1} for a given state x_t and given action a_t.

In a Markov decision problem, future states of the system 201 are thus not a function of states and actions which lie further in the past than one time step.

The characteristics of a Markov decision problem are represented below by way of summary:

X	set of possible states of the system, e.g. $X = \Re^m$,
$A(x_t)$	set of possible actions in the state
$p(x_{t+1}\|x_t, a_t)$	x_t
$r(x_t, a_t, x_{t+1})$	gain with expectation $R(x_t, a_t)$.

Starting from observable variables, the variables denoted below as training data, the aim is to determine a strategy, that is to say a sequence of functions

$$\pi = \{\mu_0, \mu_1, K, \mu_T\}, \tag{3}$$

which at each instant t map each state into an action rule, that is to say action

$$\mu_t(x_t) = a_t \tag{4}$$

Such a strategy is evaluated by an optimization function.

The optimization function specifies the expectation, the gains accumulated over time at a given strategy π, and a start state x_0.

The so-called Q-learning method is described in [1] as an example of a method of approximative dynamic programming.

An optimum evaluation function $V^*(x)$ is defined by

$$V^*(x) = \max_{\pi} V^{\pi}(x) \qquad \forall x \in X \tag{5}$$

where

$$V^{\pi}(x) = E\left[\sum_{t=0}^{\infty} \gamma^t r(x_t, \mu_t, x_{t+1}) | x_0 = x\right], \tag{6}$$

γ denoting a prescribable reduction factor which is formed in accordance with the following rule:

$$\gamma = \frac{1}{1+z}, \tag{7}$$

$$z \in \Re^+. \tag{8}$$

A Q-evaluation function $Q^*(x_t, a_t)$ is formed within the Q-learning method for each pair (state x_t, action a_t) in accordance with the following rule:

$$\begin{aligned} Q^*(x_t, a_t) = \; & \sum_{x \in X} p(x_{t+1} | x_t, a_t) \cdot r_t + \\ & + \gamma \cdot \sum_{x \in X} p(x | x_t, a_t) \cdot \max_{a \in A}\left(Q^*(x, a)\right) \end{aligned}$$

(9)

On the basis respectively of the tupel (x_t, x_{t+1}, a_t, r_t), the Q-values Q* (x,a) are adapted in the k+1 th iteration in accordance with the following learning rule with a prescribed learning rate η_k in accordance with the following rule:

$$Q_{k+1}\left(x_t, a_t\right) = \left(1 - \eta_k\right)Q_k\left(x_t, a_t\right) + \eta_k\left(r_t + \gamma \max_{a \in A}\left(Q_k\left(x_{t+1}, a\right)\right)\right). \quad (10)$$

Usually, the so-called Q-values Q*(x,a) are approximated for various actions a by a function approximator in each case, for example a neural network or else a polynomial classifier, with a weighting vector w^a, which contains weights of the function approximator.

A function approximator is to be understood as, for example, a neural network, a polynomial classifier or else a combination of a neural network with a polynomial classifier.

It therefore holds that:

$$Q^*(x, a) \approx Q\left(x; w^a\right). \quad (11)$$

Changes in the weights in the weighting vector w^a are based on a temporal difference d_t which is formed in accordance with the following rule:

$$d_t := r\left(x_t, a_t, x_{t+1}\right) + \gamma \max_{a \in A} Q\left(x_{t+1}; w_k^a\right) - Q\left(x_t; w_k^{a_t}\right) \quad (12)$$

The following adaptation rule for the weights of the neural network, which are included in the weighting vector w^a, follows for the Q-learning method with the use of a neural network:

$$w_{k+1}^{a_t} = w_k^{a_t} + \eta_k \cdot d_t \cdot \nabla Q\left(x_t; w_k^{a_t}\right). \tag{13}$$

The neural network representing the system of a financial market, as described in [1], is trained using the training data which describe information on changes in prices on a financial market as time series values.

A further method of approximative dynamic programming, the so-called TD(λ)-learning method, is known from [2] and is explained in more detail in conjunction with an exemplary embodiment.

Furthermore, it is known from [3] which risk is associated with a strategy π and an initial state x_t. A method for risk avoidment is likewise known from [3].

The following optimization function, which is also referred to as an expanded Q-function $\underline{Q}^{\pi}(x_t, a_t)$, is used in the method known from [3]:

$$\text{maximize}\left(\underline{Q}^{\pi}(x_t, a_t) := r(x_t, a_t, x_t+1) + \inf_{\substack{x_0, x_1, K \\ p(x_0, x_1, K) > 0}} \left\{\sum_{k=1}^{\infty} \gamma^k r\left(x_k, \pi(x_k), x_{k+1}\right)\right\}\right) \tag{14}$$

The expanded Q-function $\underline{Q}^{\pi}(x_t, a_t)$ describes the worst case if the action a_t is executed in the state x_t and the strategy π is followed thereupon.

The optimization function $\underline{Q}^{\pi}(x_t, a_t)$ for

$$\underline{Q}^{*}(x_t, a_t) := \max_{\pi \in \Pi} \underline{Q}^{\pi}(x_t, a_t) \qquad (15)$$

,

is given by the following rule:

$$\underline{Q}^{*}(x_t, a_t) = \min_{\substack{x \in X \\ p(x_{t+1}|x_t, a_t) > 0}} \left(r(x_t, a_t, x) + \gamma \cdot \max_{a \in A} \underline{Q}^{*}(x, a) \right). \qquad (16)$$

A substantial disadvantage of this mode of procedure is to be seen in that only the worst case is taken into account when finding the strategy. However, this reflects the requirements of the most varied technical systems only to an inadequate extent.

Furthermore, it is known from [4] to formulate access control for a communications network and the routing within the communications network as a problem of dynamic programming.

The invention is therefore based on the problem of specifying a method and an arrangement for determining a sequence of actions for a system, in which method or action an increased flexibility in determining the strategy is achieved.

The problem is solved by the method and by the arrangement in accordance with the features of the independent patent claims.

In a method for computer-aided determination of a sequence of actions for a system which has states, a transition in state between two states being performed on the basis of an action, the determination of the sequence of actions is performed in such a way that a sequence of states resulting from the sequence of actions is optimized with regard to a prescribed optimization function, the optimization function including a variable parameter with the aid of which it is possible to set a risk which the resulting sequence of states has with respect to a prescribed state of the system.

An arrangement for determining a sequence of actions for a system which has states, a transition in state between two states being performed on the basis of an action, has a processor which is set up in such a way that the determination of the sequence of actions can be performed in such a way that a sequence of states resulting from the sequence of actions is optimized with regard to a prescribed optimization function, the optimization function including a variable parameter with the aid of which it is possible to set a risk which the resulting sequence of states has with respect to a prescribed state of the system.

It becomes possible for the first time owing to the invention to specify a method for determining a sequence of actions at a freely prescribable level of accuracy when finding a strategy for a possible closed-loop control or open-loop control of the system, in general for influencing it.

Preferred developments of the invention follow from the dependent claims.

The developments described below are valid both for the method and for the arrangement, the processor being respectively set up in the development of arrangement in such a way that the development can be implemented.

In a preferred refinement, a method of approximative dynamic programming is used for the purpose of determination, for example a method based on Q-learning or else a method based on TD(λ)-learning.

Within Q-learning, the optimization function OFQ is preferably formed in accordance with the following rule:

$$OFQ = Q\left(x; w^a\right),$$

- x denoting a state in a state space X
- a denoting an action from an action space A, and
- w^a denoting the weights of a function approximator which belong to the action a.

The following adaptation step is executed during Q-learning in order to determine the optimum weights w^a of the function approximator:

$$w_{t+1}^{a_t} = w_t^{a_t} + \eta_t \cdot \aleph^{\kappa}\left(d_t\right) \cdot \nabla Q\left(x_t; w_t^{a_t}\right)$$

with the abbreviation

$$d_t = r\left(x_t, a_t, x_{t+1}\right) + \gamma \max_{a \in A} Q\left(x_{t+1}, w_t^a\right) - Q\left(x_t, w_t^{a_t}\right)$$

- x_t, x_t+1 respectively denoting a state in the state space X,
- a_t denoting an action from an action space A,
- γ denoting a prescribable reduction factor,
- $w_t^{a_t}$ denoting the weighting vector associated with the action a_t before the adaptation step,
- $w_{t+1}^{a_t}$ denoting the weighing vector associated with the action a_t after the adaptation step,
- η_t (t = 1, ...) denoting a prescribable step size sequence,

- $\kappa \in [-1; 1]$ denoting a risk monitoring parameter,
- $\aleph^\kappa$ denoting a risk monitoring function $\aleph^\kappa(\xi) = (1 - \kappa\mathrm{sign}(\xi))\xi$,
- $\nabla \mathbf{Q}(\cdot;\cdot)$ denoting the derivation of the function approximator according to its weights, and
- $r(x_t, a_t, x_{t+1})$ denoting a gain upon the transition of state from the state x_t to the subsequent state x_{t+1}.

The optimization function is preferably formed in accordance with the following rule within the TD(λ)-learning method:

$$OFTD = J(x;w)$$

- x denoting a state in a state space X,
- a denoting an action from an action space A, and
- w denoting the weights of a function approximator.

The following adaptation step is executed during TD(λ)-learning in order to determine the optimum weights w of the function approximator:

$$w_{t+1} = w_t + \eta_t \cdot \aleph^\kappa(d_t) \cdot z_t$$

with the abbreviations

$$d_t = r(w_t, a_t, x_{t+1}) + \gamma J(x_{t+1}; w_t) - J(x_t; w_t),$$

$$z_t = \lambda \cdot \gamma \cdot z_{t-1} + \nabla J(x_t; w_t),$$

$$z_{-1} = 0$$

- x_t, x_{t+1} respectively denoting a state in the state space X,
- a_t denoting an action from an action space A,
- γ denoting a prescribable reduction factor,
- w_t denoting the weighting vector before the adaptation step,
- w_{t+1} denoting the weighting vector after the adaptation step,
- η_t (t = 1, ...) denoting a prescribable step size sequence,
- $\kappa \in [-1; 1]$ denoting a risk monitoring parameter,
- $\aleph^\kappa$ denoting a risk monitoring function $\aleph^\kappa(\xi) = (1 - \kappa\mathrm{sign}(\xi))\xi$,
- $\nabla J(\cdot;\cdot)$ denoting the derivation of the function approximator according to its weights, and
- $r(x_t, a_t, x_{t+1})$ denoting a gain upon the transition of state from the state x_t to the subsequent state x_{t+1}.

The system is preferably a technical system of which before the determination measured values are measured which are used in determining the sequence of actions.

The technical system can be subjected to open-loop control or else closed-loop control with the use of the determined sequence of actions.

The system is preferably modeled as a Markov decision problem.

The method or the arrangement is preferably used in a traffic management system or in a communications system, the sequence of actions being used in a communications network to carry out access control or a routing, that is to say a path allocation.

Furthermore, the system can be a financial market which is modeled by a Markov decision problem, the change in the financial market, for example the change in an index of stocks or else a rate of exchange on a foreign exchange market being analyzed by using the method and/or the arrangement and it being possible to intervene in the market in accordance with the sequence of determined actions.

Exemplary embodiments of the invention are illustrated in the figures and explained in more detail below.

Figure 1 shows a flowchart in which individual method steps of the first exemplary embodiment are illustrated;

Figure 2 shows a sketch of a system which can be modeled as a Markov decision problem;

Figure 3 shows a sketch of a communications network in which access control is carried out in a switching unit;

Figure 4 shows a symbolic sketch of a function approximator with the aid of which a method of approximative dynamic programming is implemented;

Figure 5 shows a further sketch of a plurality of function approximators with the aid of which approximative dynamic programming is implemented; and

Figure 6 shows a sketch of a traffic management system which is subjected to closed-loop control in accordance with an exemplary embodiment.

First exemplary embodiment: access control and routing.

Figure 3 shows a communications network 300, which has a multiplicity of switching units 301a, 301b, ..., 301i, ... 301n, which are interconnected via connections 302a, 302b, 302j, ... 302m.

Furthermore, a first terminal 303 is connected to a first switching unit 301a. From the first terminal 303, the first switching unit 301a is sent a request message 304 which requests preservation of a prescribed bandwidth within the communications network 300 for the purpose of transmitting data (video data, text data).

It is determined in the first switching unit 301a in accordance with a strategy described below whether the requested bandwidth is available in the communications network 300 on a specified, requested connection (step 305).

The request is refused (step 306) if this is not the case.

If sufficient bandwidth is available, it is checked in a further checking step (step 307) whether the bandwidth can be reserved.

The request is refused (step 308) if this is not the case.

Otherwise, the first switching unit 301a selects a route from the first switching unit 301a via further switching units 301i to a second terminal 309 with which the first terminal 303 wishes to communicate, and a connection is initialized (step 310).

The starting point below is a communications network 300 which comprises a set of switching units

$$N = \{1, K, n, K, N\} \qquad (17)$$

and a set of physical connections

$$L = \{1, K, 1, K, L\}, \qquad (18)$$

a physical connection 1 having a capacity of B(1) bandwidth units.

A set

$$M = \{1, K, m, K, M\} \qquad (19)$$

of different types of service m are available, a type of service m being characterized by

- a bandwidth requirement b(m),
- an average connection time $\frac{1}{V(m)}$ and
- a gain c(m) which is obtained whenever a call request of the corresponding type of service m is accepted.

The gain c(m) is given by the amount of money which a network operator of the communications network 300 bills a subscriber for a connection of the type of service. Clearly, the gain c(m) reflects different priorities, which can be prescribed by the network operator and which he associates with different services.

A physical connection 1 can simultaneously provide any desired combination of communications connections as long as the bandwidth used for the communications connections does not exceed the bandwidth available overall for the physical connection.

If a new communications connection of type m is requested between a first node i and a second node j (terminals are also denoted as nodes), the requested communications connection can, as represented above, either be accepted or be refused.

If the communications connection is accepted, a route is selected from a set of prescribed routes. This selection is denoted as a routing. b(m) bandwidth units are used in the communications connection of type m for each physical connection along the selected route for the duration of the connection.

Thus, during access control (call admission control), a route can be selected within the communications network 300 only when the selected route has sufficient bandwidth available.

The aim of the access control and of the routing is to maximize a long term gain which is obtained by acceptance of the requested connections.

At an instant t, the technical system which is the communications network 300 is in a state x_t which is described by a list of routes via existing connections, by means of which lists it is shown how many connections of which type of service are using the respective routes at the instant t.

Events w, by means of which a state x_t could be transferred into a subsequent state x_{t+1}, are the arrival of new connection request messages, or else the termination of a connection existing in the communications network 300.

In this exemplary embodiment, an action a_t at an instant t owing to a connection request is the decision as to whether a connection request is to be accepted or refused and, if the connection is accepted, the selection of the route through the communications network 300.

The aim is to determine a sequence of actions, that is to say clearly to determine the learning of a strategy with actions relating to a state x_t in such a way that the following rule is maximized:

$$E\left(\sum_{k=0}^{\infty} e^{-\beta t_k} \cdot g\left(x_{t_k}, \omega_k, a_{t_k}\right)\right), \tag{20}$$

- $E\{.\}$ denoting an expectation,
- t_k denoting an instant at which a kth event takes place,
- $g\left(x_{t_k}, \omega_k, a_{t_k}\right)$. denoting the gain which is associated with the kth event, and
- β denoting a reduction factor which evaluates an immediate gain as being more valuable than a gain at instants lying further in the future.

Different implementations of a strategy lead normally to different overall gains G:

$$G = \sum_{k=0}^{\infty} e^{-\beta t_k} \cdot g\left(x_{t_k}, \omega_k, a_{t_k}\right). \tag{21}$$

The aim is to maximize the expectation of the overall gain G in accordance with the following rule J:

$$J = E\left\{\sum_{k=0}^{\infty} e^{-\beta t_k} \cdot g\left(x_{t_k}, \omega_k, a_{t_k}\right)\right\}, \tag{22}$$

it being possible to set a risk which reduces the overall gain G of a specific implementation of access control and of a routing strategy to below the expectation.

The TD(λ)-learning method is used to carry out the access control and the routing.

The following target function is used in this exemplary embodiment:

$$J^*(x_t) = E_\tau\left\{e^{-\beta\tau}\right\}E_\omega\left\{\max_{a \in A}\left[g(x_t, \omega_t, a) + J^*(x_{t+1})\right]\right\}, \tag{23}$$

- A denoting an action space with a prescribed number of actions which are respectively available in a state x_t,
- τ denoting a first instant at which a first event ω occurs, and
- x_{t+1} denoting a subsequent state of the system.

An approximated value of the target value $J^*(x_t)$ is learned and stored by employing a function approximator 400 (compare **Figure 4**) with the use of training data.

Training data are data previously measured in the communications network 300 and relating to the behavior of the communications network 300 in the case of incoming connection requests 304 and of termination of messages. This time sequence of states is stored, and these training data are used to train the function approximator 400 in accordance with the learning method described below.

A number of connections of in each case one type of service m on a route of the communications network 300 serve in each case as input variable of the function approximator 400 for each input 401, 402, 403 of the function approximator 400. These are represented symbolically in **Figure 4** by blocks 404, 405, 406.

An approximated target value $\tilde{J}$ of the target value J^* is the output variable of the function approximator 400.

Figure 5 shows a detailed representation of the function approximator 500, which in this case has several component function approximators 510, 520 of the function approximator 500. One output variable is the approximated target value $\tilde{J}$, which is formed in accordance with the following rule:

$$\tilde{J}(x_t, \Theta) = \sum_{l=1}^{L} \tilde{J}^{(l)}\left(x_t^{(l)}, \Theta_t^{(l)}\right). \qquad (24)$$

The input variables of the component function approximators 510, 520, which are present at the inputs 511, 512, 513 of the first component function approximator 510, or at the inputs 521, 522 and 523 of the second component function approximator 520 are, in turn, respectively a number of types of service of a type m in a physical connection r in each case, symbolized by blocks 514, 515, 516 for the first component function approximator, and 524, 525 and 526 for the second component function approximator 520.

Component output variables 530, 531, 532, 533 are fed to an adder unit 540, and the approximated target variable $\tilde{J}$ is formed as output variable of the adder unit.

Let it be assumed that the communications network 300 is in the state x_{tk} and that a request message with which a type of service m of class m is requested for a connection between two nodes i, j reaches the first switching unit 301a.

A list of permitted routes between the nodes i and j is denoted by R(i, j), and a list of all possible routes is denoted by

$$\tilde{R}\left(i, j, x_{t_k}\right) \subset R(i, j) \qquad (25)$$

as a subset of the routes R(i, j) which could implement a possible connection with regard to the available and requested bandwidth.

For each possible route r, $r \in \tilde{R}\left(i, j, x_{t_k}\right)$, a subsequent state $x_{t_k+1}\left(x_{t_k}, \omega_k, r\right)$ is determined which results from the fact that the connection request 304 is accepted and the connection on the route r is made available to the requesting first terminal 303.

This is illustrated in **Figure 1** as second step (step 102), the state of the system and the respective event being respectively determined in a first step (step 101).

A route r* to be selected is determined in a third step (step 103) in accordance with the following rule:

$$r^* = \arg \max_{r \in \tilde{R}\left(i, j, x_{t_k}\right)} \tilde{J}\left(x_{t_k+1}\left(x_{t_k}, \omega_k, r\right), \Theta_t\right). \qquad (26)$$

A check is made in a further step (step 104) as to whether the following rule is fulfilled:

$$c(m) + \tilde{J}\left(x_{t_k+1}\left(x_{t_k}, \omega_k, r^*\right), \Theta_t\right) < \tilde{J}\left(x_{t_k}, \Theta_t\right). \qquad (27)$$

If this is the case, the connection request 304 is rejected (step 105), otherwise the connection is accepted and "switched through" to the node j along the selected route r* (step 106).

Weights of the function approximator 400, 500 which are adapted in the TD(λ)-learning method to the training data, are stored in a parameter vector θ for an instant t in each case, such that an optimized access control and an optimized routing are achieved.

During the training phase, the weighting parameters are adapted to the training data applied to the function approximator.

A risk parameter κ is defined with the aid of which a desired risk, which the system has with regard to a prescribed state owing to a sequence of actions and states, can be set in accordance with the following rules:

$-1 \leq \kappa < 0$:	risky learning,
$\kappa = 0$:	neutral learning with regard to the risk,
$0 < \kappa < 1$:	risk-avoiding learning,
$\kappa = 1$:	worst-case learning.

Furthermore, a prescribable parameter $0 \leq \lambda \leq 1$ and a step size sequence γ_k are prescribed in the learning method.

The weighting values of the weighting vector Θ are adapted to the training data on the basis of each event ω_{t_k} in accordance with the following adaptation rule:

$$\Theta_k = \Theta_{k-1} + \gamma_k \aleph^{\kappa}(d_k) z_t , \qquad (28)$$

in which case

$$d_k = e^{-\beta(t_k - t_{k-1})}\left(g\left(x_{t_k}, \omega_k, a_{t_k}\right) + \tilde{J}\left(x_{t_k}, \Theta_{k-1}\right)\right) - \tilde{J}\left(x_{t_{k-1}}, \Theta_{k-1}\right) \tag{29}$$

$$z_t = \lambda e^{-\beta(t_{k-1} - t_{k-2})} z_{t-1} + \nabla_{\Theta}\tilde{J}\left(x_{t_{k-1}}, \Theta_{k-1}\right), \tag{30}$$

and

$$\aleph^{\kappa}(\xi) = \left(1 - \kappa \operatorname{sign}(\xi)\right)\xi\,. \tag{31}$$

It is assumed that: $z_{-1} = 0$.

The function

$$g\left(x_{t_k}, \omega_k, a_{t_k}\right) \tag{32}$$

denotes the immediate gain in accordance with the following rule:

$$g\left(x_{t_k}, \omega_k, a_{t_k}\right) = \begin{cases} c(m) & \text{when } \omega_{t_k} \text{ is a service request for a type of service m, and the connection is accepted} \\ 0 & \text{otherwise} \end{cases} \tag{33}$$

Thus, as described above, a sequence of actions is determined with regard to a connection request such that a connection request is either rejected or accepted on the basis of an action. The determination is performed taking account of an optimization function in which the risk can be set by means of a risk control parameter $\kappa \in [-1; 1]$ in a variable fashion.

Second exemplary embodiment: Traffic management system

Figure 6 shows a road 600 on which automobiles 601, 602, 603, 604, 605 and 606 are being driven.

Conductor loops 610, 611 integrated into the road 600 receive electric signals in a known way and feed the electric signals 615, 616 to a computer 620 via an input/output interface 621. In an analog-to-digital converter 622 connected to the input/output interface 621, the electric signals are digitized into a time series and stored in a memory 623, which is connected by a bus 624 to the analog-to-digital converter 622 and a processor 625. Via the input/output interface 621, a traffic management system 650 is fed control signals 651 from which it is possible to set a prescribed speed stipulation 652 in the traffic management system 650, or else further particulars of traffic regulations, which are displayed via the traffic management system 650 to drivers of the vehicles 601, 602, 603, 604, 605 and 606.

The following local state variables are used in this case for the purpose of traffic modeling:

- traffic flow rate v,
- vehicle density ρ (ρ = number of vehicles per kilometer $\frac{Fz}{km}$).
- traffic flow q (q = number of vehicles per hour $\frac{Fz}{h}$, (q= v * ρ)), and
- speed restrictions 652 displayed by the traffic management system 650 at an instant in each case.

The local state variables are measured as described above by using the conductor loops 610, 611.

These variables $(v(t), \rho(t), q(t))$ therefore represent a state of the technical system of "traffic" at a specific instant t.

In this exemplary embodiment, the system is therefore a traffic system which is controlled by using the traffic management system 650.

In this second exemplary embodiment, an extended Q-learning method is described as method of approximative dynamic programming.

The state x_t is described by a state vector

$$x(t) := (v(t), \rho(t), q(t)). \tag{34}$$

The action a_t denotes the speed restriction 652, which is displayed at the instant t by the traffic management system 650.

The gain $r(x_t, a_t, x_{t+1})$ describes the quality of the traffic flow which was measured between the instants t and t+1 by the conductor loops 610 and 611. In this second exemplary embodiment, $r(x_t, a_t, x_{t+1})$ denotes

- the average speed of the vehicles in the time interval [t, t + 1]

or

- the number of vehicles which have passed the conductor loops 610 and 611 in the time interval [t, t + 1]

or

- the variance of the vehicle speeds in the time interval [t, t + 1], or

- a weighted sum from the above variables.

A value of the optimization function OFQ is determined for each possible action a_t, that is to say for each speed restriction which can be displayed by the traffic management system 650, an estimated value of the optimization function OFQ being realized in each case as a neural network.

This results in a set of evaluation variables for the various actions a_t in the system state x_t.

Those actions a_t for which the maximum evaluation variable OFQ has been determined in the current system state x_t are selected in a control phase from the possible actions a_t, that is to say from the set of the speed restrictions which can be displayed by the traffic management system 650.

In accordance with this exemplary embodiment, the adaptation rule, known from the Q-learning method, for calculating the optimization function OFQ is extended by a risk control function $\aleph^\kappa(.)$, which takes account of the risk.

In turn, the risk control parameter κ is prescribed in accordance with the strategy from the first exemplary embodiment in the interval of $[-1 \le \kappa \le 1]$, and represents the risk which a user wishes to run in the application with regard to the control strategy to be determined.

The following evaluation function OFQ is used in accordance with this exemplary embodiment:

$$OFQ = Q\left(x; w^a\right), \tag{35}$$

- $x = (v;\ \rho;\ q)$ denoting a state of the traffic system,
- a denoting a speed restriction from the action space A of all speed restrictions which can be displayed by the traffic management system 650, and
- w^a denoting the weights of the neural network which belong to the speed restriction a.

The following adaptation step is executed in Q-learning in order to determine the optimum weights w^a of the neural network:

$$w_{t+1}^{a_t} = w_t^{a_t} + \eta_t \cdot \aleph^{\kappa}(d_t) \cdot \nabla Q\left(x_t; w_t^{a_t}\right) \qquad (36)$$

using the abbreviation

$$d_t = r(x_t, a_t, x_{t+1}) + \gamma \max_{a \in A} Q\left(x_{t+1}, w_t^a\right) - Q\left(x_t, w_t^{a_t}\right) \qquad (37)$$

- x_t, x_{t+1} denoting in each case a state of the traffic system in accordance with rule (34),
- a_t denoting an action, that is to say a speed restriction which can be displayed by the traffic management system 650,
- γ denoting a prescribable reduction factor,
- $w_t^{a_t}$ denoting the weighting vector belonging to the action a_t, before the adaptation step,
- $w_{t+1}^{a_t}$ denoting the weighting vector belonging to the action a_t, after the adaptation step,
- η_t $(t = 1, \ldots)$ denoting a prescribable step size sequence,

- $\kappa \in [-1; 1]$ denoting a risk control parameter,
- $\aleph^{\kappa}$ denoting a risk control function $\aleph^{\kappa}(\xi) = (1 - \kappa \mathrm{sign}(\xi))\xi$,
- $\nabla_Q(\cdot;\cdot)$ denoting the derivative of the neural network with respect to its weights, and
- $r(x_t, a_t, x_{t+1})$ denoting a gain upon the transition in state from the state x_t to the subsequent state x_{t+1}.

An action a_t can be selected at random from the possible actions a_t during learning. It is not necessary in this case to select the action a_t which has led to the largest evaluation variable.

The adaptation of the weights has to be performed in such a way that not only is a traffic control achieved which is optimized in terms of the expectation of the optimization function, but that also account is taken of a variance of the control results.

This is particularly advantageous since the state vector x(t) models the actual system of traffic only inadequately in some aspects, and so unexpected disturbances can thereby occur. Thus, the dynamics of the traffic, and therefore of its modeling, depend on further factors such as weather, proportion of trucks on the road, proportion of mobile homes, etc., which are not always integrated in the measured variables of the state vector x(t). In addition, it is not always ensured that the road users immediately implement the new speed instructions in accordance with the traffic management system.

A control phase on the real system in accordance with the traffic management system takes place in accordance with the following steps:

1. The state x_t is measured at the instant t at various points in the traffic system of traffic and yields a state vector $x(t) := (v(t), \rho(t), q(t))$.

2. A value of the optimization function is determined for all possible actions a_t, and that action a_t with the highest evaluation in the optimization function is selected.

The following publications are cited in this document:

[1] R, Neuneier, Enhancing Q-Learning for Optimal Asset Allocation, Proceedings of the Neural Information Processing Systems, NIPS 1997

[2] R.S. Sutton, Learning to predict by the method of temporal differences, Machine Learning, 3:9-44, 1988

[3] M. Heger, Risk and Reinforcement Learning: Concepts and Dynamic Programming, ZKW Bericht Nr. 8/94, Zentrum für Kognitionswissenschaften [Center for Cognitive Sciences], Bremen University, ISSN 0947-0204, December 1994

[4] D.P. Bertsekas, Dynamic Programming and Optimal Control, Athena Scientific, Belmont, MA, 1995

Patent claims

1. A method for computer-aided determination of a sequence of actions for a system which has states, a transition in state between two states being performed on the basis of an action, in the case of which the determination of the sequence of actions is performed in such a way that a sequence of states resulting from the sequence of actions is optimized with regard to a prescribed optimization function, the optimization function including a variable parameter with the aid of which it is possible to set a risk which the resulting sequence of states has with respect to a prescribed state of the system.

2. The method as claimed in claim 1, in which a method of approximative dynamic programming is used for the purpose of determination.

3. The method as claimed in claim 2, in which the method of approximative dynamic programming is a method based on Q-learning.

4. The method as claimed in claim 3, in which the optimization function OFQ is formed within Q-learning in accordance with the following rule:

$$OFQ = Q\left(x; w^a\right),$$

- x denoting a state in a state space X
- a denoting an action from an action space A, and
- w^a denoting the weights of a function approximator which belong to the action a,

and in which the weights of the function approximator are adapted in accordance with the following rule:

$$w_{t+1}^{a_t} = w_t^{a_t} + \eta_t \cdot \aleph^{\kappa}(d_t) \cdot \nabla Q\left(x_t; w_t^{a_t}\right)$$

with the abbreviation

$$d_t = r\left(x_t, a_t, x_{t+1}\right) + \gamma \max_{a \in A} Q\left(x_{t+1}, w_t^{a}\right) - Q\left(x_t, w_t^{a_t}\right)$$

- x_t, x_t+1 respectively denoting a state in the state space X,
- a_t denoting an action from an action space A,
- γ denoting a prescribable reduction factor,
- $w_t^{a_t}$ denoting the weighting vector associated with the action a_t before the adaptation step,
- $w_{t+1}^{a_t}$ denoting the weighing vector associated with the action a_t after the adaptation step,
- η_t (t = 1, ...) denoting a prescribable step size sequence,
- $\kappa \in [-1; 1]$ denoting a risk monitoring parameter,
- $\aleph^{\kappa}$ denoting a risk monitoring function $\aleph^{\kappa}(\xi) = (1 - \kappa \mathrm{sign}(\xi))\xi$,
- $\nabla Q(\cdot;\cdot)$ denoting the derivation of the function approximator according to its weights, and
- $r(x_t, a_t, x_{t+1})$ denoting a gain upon the transition of state from the state x_t to the subsequent state x_{t+1}.

5. The method as claimed in claim 2, in which the method of approximative dynamic programming is a method based on TD(λ)-learning.

6. The method as claimed in claim 5, in which the optimization function OFTD is formed within TD(λ)-learning in accordance with the following rule:

OFTD = $J(x;w)$

- x denoting a state in a state space X,
- a denoting an action from an action space A, and
- w denoting the weights of a function approximator

and in which the weights of the function approximator are adapted in accordance with the following rule:

$$w_{t+1} = w_t + \eta_t \cdot \aleph^{\kappa}(d_t) \cdot z_t$$

with the abbreviations

$$d_t = r(w_t,\ a_t,\ x_{t+1}) + \gamma J(x_{t+1};\ w_t) - J(x_t;\ w_t),$$

$$z_t = \lambda \cdot \gamma \cdot z_{t-1} + \nabla J(x_t;\ w_t),$$

$$z_{-1} = 0$$

- x_t, x_{t+1} respectively denoting a state in the state space X,
- a_t denoting an action from an action space A,
- γ denoting a prescribable reduction factor,
- w_t denoting the weighting vector before the adaptation step,
- w_{t+1} denoting the weighting vector after the adaptation step,
- η_t ($t = 1, \ldots$) denoting a prescribable step size sequence,
- $\kappa \in [-1; 1]$ denoting a risk monitoring parameter,
- $\aleph^{\kappa}$ denoting a risk monitoring function $\aleph^{\kappa}(\xi) = (1 - \kappa \mathrm{sign}(\xi))\xi$,
- $\nabla J(\cdot;\cdot)$ denoting the derivation of the function approximator according to its weights, and
- $r(x_t,\ a_t,\ x_{t+1})$ denoting a gain upon the transition of state from the state x_t to the subsequent state x_{t+1}.

7. The method as claimed in one of claims 1 to 6, in which the system is a technical system of which before the determination measured values are measured which are used in determining the sequence of actions.

8. The method as claimed in claim 7, in which the technical system is subjected to open-loop control in accordance with the sequence of actions.

9. The method as claimed in claim 7, in which the technical system is subjected to closed-loop control in accordance with the sequence of actions.

10. The method as claimed in one of claims 1 to 9, in which the system is modeled as a Markov decision problem.

11. The method as claimed in one of claims 1 to 10, being used in a traffic management system.

12. The method as claimed in one of claims 1 to 10, being used in a communications system.

13. The method as claimed in one of claims 1 to 10, being used to carry out access control in a communications network.

14. The method as claimed in one of claims 1 to 10, being used to carry out a routing in a communications network.

15. An arrangement for determining a sequence of actions for a system which has states, a transition in state between two states being performed on the basis of an action,

having a processor which is set up in such a way that the determination of the sequence of actions can be performed in such a way that a sequence of states resulting from the sequence of actions is optimized with regard to a prescribed optimization function, the optimization function including a variable parameter with the aid of which it is possible to set a risk which the resulting sequence of states has with respect to a prescribed state of the system.

16. The arrangement as claimed in claim 15, being used to subject a technical system to open-loop control.

17. The arrangement as claimed in claim 15, being used to subject a technical system to closed-loop control.

18. The arrangement as claimed in claim 15, being used in a traffic management system.

19. The arrangement as claimed in claim 15, being used in a communications system.

20. The arrangement as claimed in claim 15, being used to carry out access control in a communications network.

21. The arrangement as claimed in claim 15, being used to carry out a routing in a communications network.

GR 98 P 2663

Abstract

Method and arrangement for determining a sequence of actions for a system which has states, a transition in state between two states being performed on the basis of an action

The determination of a sequence of actions is performed in such a way that a sequence of states resulting from the sequence of actions is optmized with regard to a prescribed optimization function. The optimization function includes a variable parameter with the aid of which it is possible to set a risk which the resulting sequence of states has with respect to a prescribed state of the system.

FIG 1

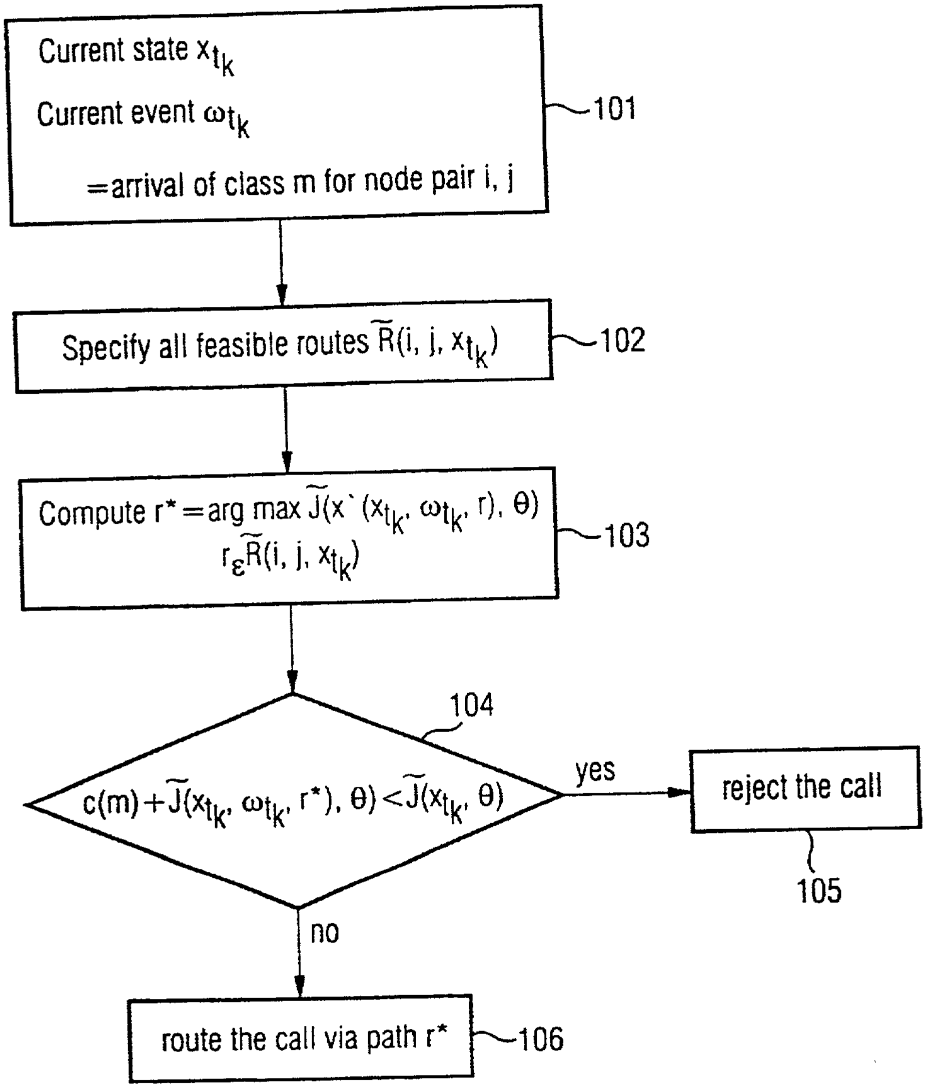

Current state x_{t_k}
Current event ω_{t_k}
$=$arrival of class m for node pair i, j
101
Specify all feasible routes $\widetilde{R}(i, j, x_{t_k})$
102
Compute $r^* = \arg\max_{r \varepsilon \widetilde{R}(i, j, x_{t_k})} \widetilde{J}(x`(x_{t_k}, \omega_{t_k}, r), \theta)$
103
104
$c(m) + \widetilde{J}(x_{t_k}, \omega_{t_k}, r^*), \theta) < \widetilde{J}(x_{t_k}, \theta)$
yes
reject the call
105
no
route the call via path r*
106

FIG 2

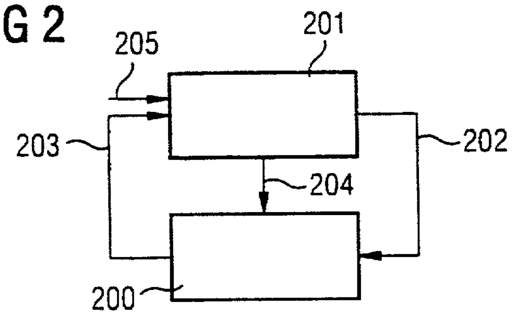

205
201
203
202
204
200

FIG3

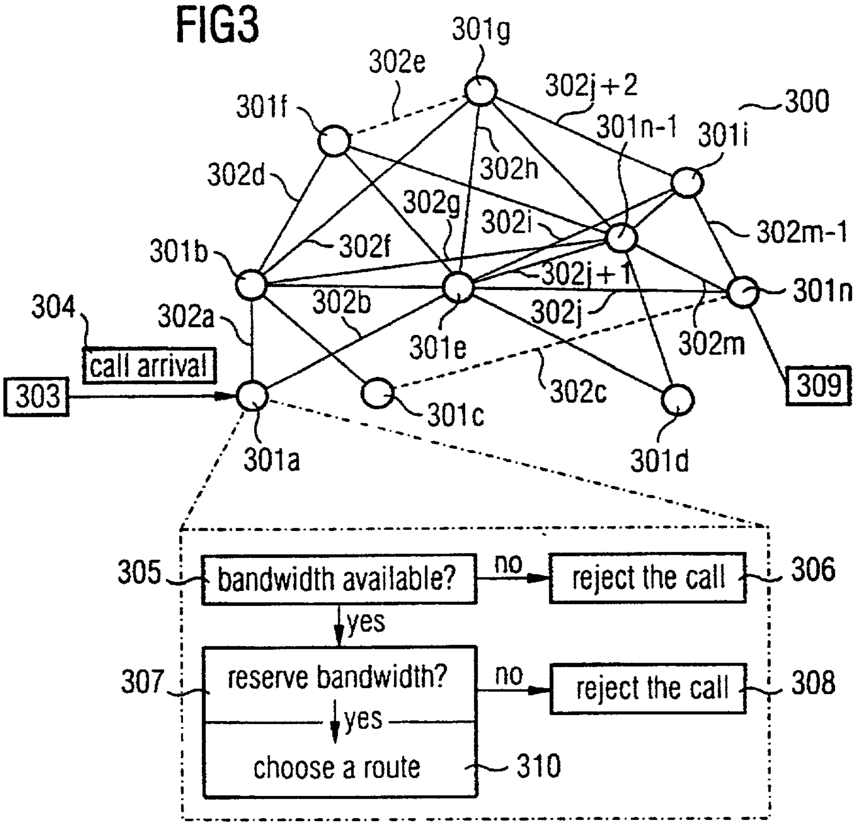

301g
302e
302j+2
300
301f
301n-1
301i
302d
302h
302g
302i
302m-1
301b
302f
302j+1
301n
304
302a
302b
302j
call arrival
301e
302m
303
302c
309
301c
301a
301d
305
bandwidth available?
no
reject the call
306
yes
307
reserve bandwidth?
no
reject the call
308
yes
choose a route
310

FIG 4

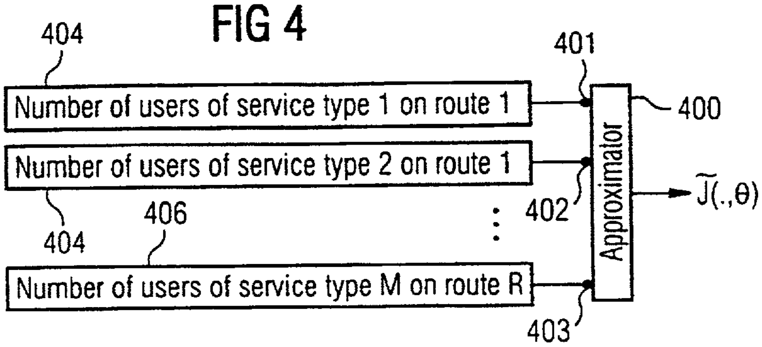

404
Number of users of service type 1 on route 1
Number of users of service type 2 on route 1
404
406
Number of users of service type M on route R
401
402
403
400
Approximator
$\tilde{J}(.,\theta)$

FIG 5

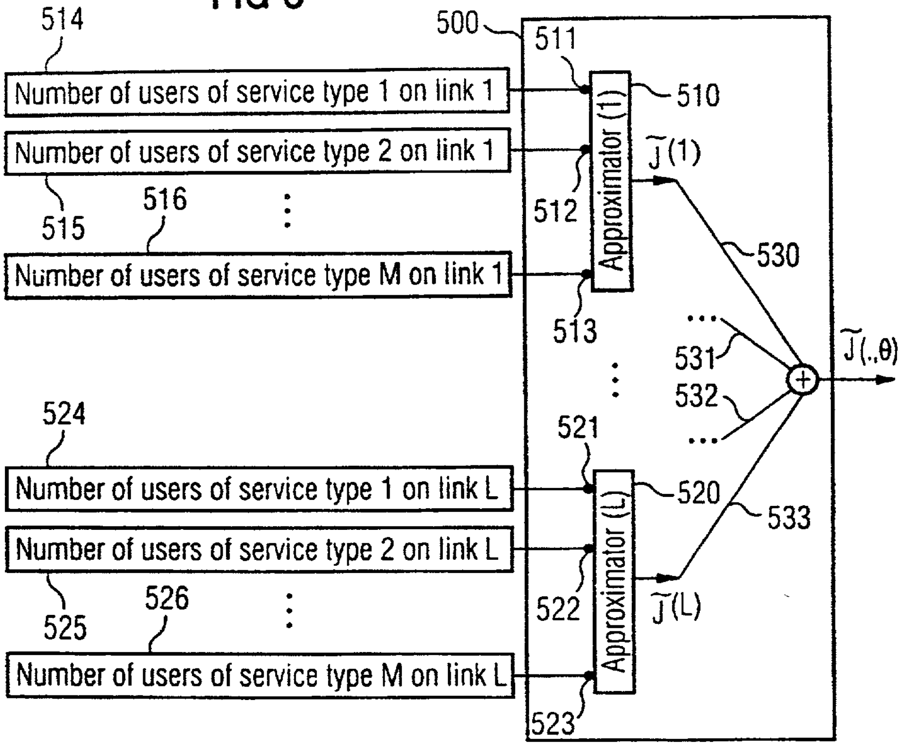
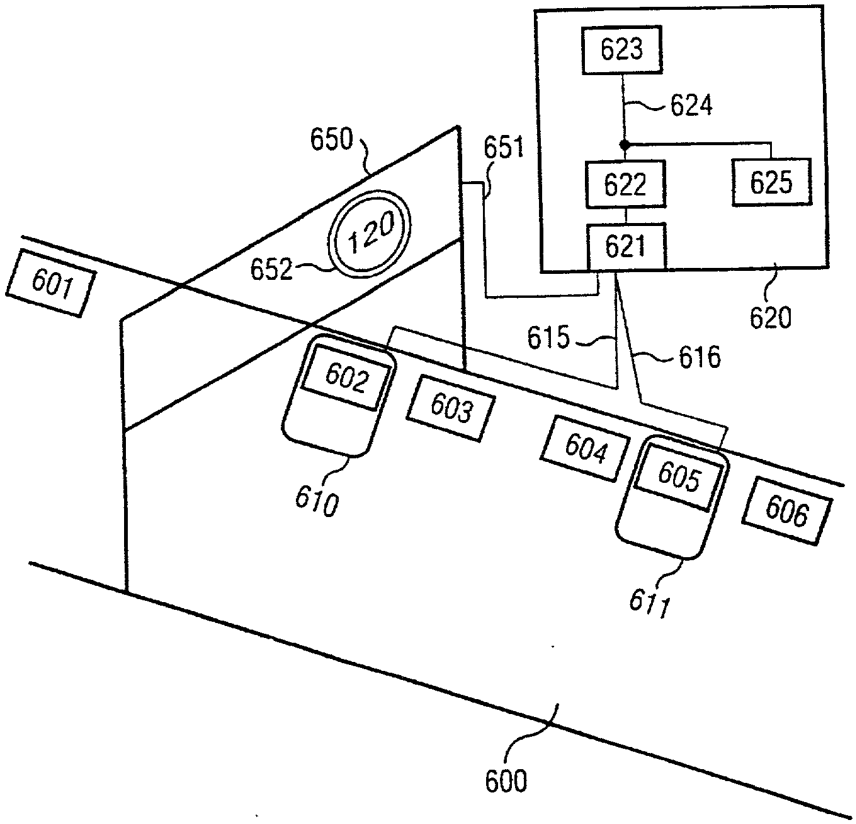

514
Number of users of service type 1 on link 1
Number of users of service type 2 on link 1
515
516
Number of users of service type M on link 1
524
Number of users of service type 1 on link L
Number of users of service type 2 on link L
525
526
Number of users of service type M on link L
500
511
512
513
510
Approximator (1)
$\tilde{J}(1)$
530
531
532
533
$\tilde{J}(.,\theta)$
521
522
523
520
Approximator (L)
$\tilde{J}(L)$

FIG 6


623
624
650
651
622
625
120
652
621
601
620
615
616
602
603
604
605
606
610
611
600

BOX PCT
IN THE UNITED STATES DESIGNATED/ELECTED OFFICE
OF THE UNITED STATES PATENT AND TRADEMARK OFFICE
UNDER THE PATENT COOPERATION TREATY-CHAPTER II

CHANGE OF ADDRESS OF APPLICANTS' REPRESENTATIVE

APPLICANT(S):	RALF NEUNEIER, ET AL.
ATTORNEY DOCKET NO.:	P01,0020
INTERNATIONAL APPLICATION NO:	PCT/DE99/02846
INTERNATIONAL FILING DATE:	8 SEP 1999
INVENTION:	METHOD AND ARRANGEMENT FOR DETERMINING A SEQUENCE OF ACTIONS FOR A SYSTEM

Assistant Commissioner for Patents,
Washington D.C. 20231

S I R:

Members of the firm of Hill & Simpson designated on the original Power of Attorney have merged into the firm of Schiff Hardin & Waite. All future correspondence for the above-referenced application therefore should be sent to the following address:

SCHIFF HARDIN & WAITE
Patent Department
6600 Sears Tower
233 South Wacker Drive
Chicago, Illinois 60606-6473

CUSTOMER NUMBER 26574

Submitted by,

Steven H. Noll (Reg. No. 28,982)
Steven H. Noll
SCHIFF HARDIN & WAITE
Patent Department
6600 Sears Tower
Chicago, Illinois 60606-6473
Telephone: (312) 258-5790
Attorneys for Applicants
CUSTOMER NUMBER 26574

Declaration and Power of Attorney For Patent Application

Erklärung Für Patentanmeldungen Mit Vollmacht

German Language Declaration

Als nachstehend benannter Erfinder erkläre ich hiermit an Eides Statt:

dass mein Wohnsitz, meine Postanschrift, und meine Staatsangehörigkeit den im Nachstehenden nach meinem Namen aufgeführten Angaben entsprechen,

dass ich, nach bestem Wissen der ursprüngliche, erste und alleinige Erfinder (falls nachstehend nur ein Name angegeben ist) oder ein ursprünglicher, erster und Miterfinder (falls nachstehend mehrere Namen aufgeführt sind) des Gegenstandes bin, für den dieser Antrag gestellt wird und für den ein Patent beantragt wird für die Erfindung mit dem Titel:

Verfahren und Anordnung zur Ermittlung einer Folge von Aktionen für ein System, welches Zustände aufweist, wobei ein Zustandsübergang zwischen zwei Zuständen aufgrund einer Aktion erfolgt

deren Beschreibung

(zutreffendjhaes ankreuzen)

☒ hier beigefügt ist.

□ am ____________ als
PCT internationale Anmeldung
PCT Anmeldungsnummer ____________
Eingereicht wurde und am ____________
Abgeändert wurde (falls tatsächlich abgeändert).

Ich bestätige hiermit, dass ich den Inhalt der obigen Patentanmeldung einschliesslich der Ansprüche durchgesehen und verstanden habe, die eventuell durch einen Zusatzantrag wie oben erwähnt abgeändert wurde.

Ich erkenne meine Pflicht zur Offenbarung irgendwelcher Informationen, die für die Prüfung der vorliegenden Anmeldung in Einklang mit Absatz 37, Bundesgesetzbuch, Paragraph 1.56(a) von Wichtigkeit sind, an.

Ich beanspruche hiermit ausländische Prioritätsvorteile gemàss Abschnitt 35 der Zivilprozessordnung der Vereinigten Staaten, Paragraph 119 aller unten angegebenen Auslandsanmeldungen für ein Patent oder eine Erfindersurkunde, und habe auch alle Auslandsanmeldungen für ein Patent oder eine Erfindersurkunde nachstehend gekennzeichnet, die ein Anmeldedatum haben, das vor dem Anmeldedatum der Anmeldung liegt, für die Priorität beansprucht wird.

As a below named inventor, I hereby declare that:

My residence, post office address and citizenship are as stated below next to my name,

I believe I am the original, first and sole inventor (if only one name is listed below) or an original, first and joint inventor (if plural names are listed below) of the subject matter which is claimed and for which a patent is sought on the invention entitled

____
____
____
____
____

the specification of which

(check one)

□ is attached hereto.

□ was filed on ____________ as
PCT international application
PCT Application No. ____________
and was amended on ____________
(if applicable)

I hereby state that I have reviewed and understand the contents of the above identified specification, including the claims as amended by any amendment referred to above.

I acknowledge the duty to disclose information which is material to the examination of this application in accordance with Title 37, Code of Federal Regulations, §1.56(a).

I hereby claim foreign priority benefits under Title 35, United States Code, §119 of any foreign application(s) for patent or inventor's certificate listed below and have also identified below any foreign application for patent or inventor's certificate having a filing date before that of the application on which priority is claimed:

German Language Declaration

Prior foreign appplications
Prioritӓt beansprucht

Priority Claimed

198 43 620.3	Germany	23.September 1998	☒	□
(Number) (Nummer)	(Country) (Land)	(Day Month Year Filed) (Tag Monat Jahr eingereicht)	Yes Ja	No Nein
			□	□
(Number) (Nummer)	(Country) (Land)	(Day Month Year Filed) (Tag Monat Jahr eingereicht)	Yes Ja	No Nein
			□	□
(Number) (Nummer)	(Country) (Land)	(Day Month Year Filed) (Tag Monat Jahr eingereicht)	Yes Ja	No Nein

Ich beanspruche hiermit gemäss Absatz 35 der Zivilprozessordnung der Vereinigten Staaten, Paragraph 120, den Vorzug aller unten aufgeführten Anmeldungen und falls der Gegenstand aus jedem Anspruch dieser Anmeldung nicht in einer früheren amerikanischen Patentanmeldung laut dem ersten Paragraphen des Absatzes 35 der Zivilprozeßordnung der Vereinigten Staaten, Paragraph 122 offenbart ist, erkenne ich gemäss Absatz 37, Bundesgesetzbuch, Paragraph 1.56(a) meine Pflicht zur Offenbarung von Informationen an, die zwischen dem Anmeldedatum der früheren Anmeldung und dem nationalen oder PCT internationalen Anmeldedatum dieser Anmeldung bekannt geworden sind.

I hereby claim the benefit under Title 35. United States Code. §120 of any United States application(s) listed below and, insofar as the subject matter of each of the claims of this application is not disclosed in the prior United States application in the manner provided by the first paragraph of Title 35, United States Code, §122, I acknowledge the duty to disclose material information as defined in Title 37, Code of Federal Regulations, §1.56(a) which occured between the filing date of the prior application and the national or PCT international filing date of this application.

(Application Serial No.) (Anmeldeseriennummer)	(Filing Date) (Anmeldedatum)	(Status) (patentiert, anhängig, aufgegeben)	(Status) (patented, pending, abandoned)
(Application Serial No.) (Anmeldeseriennummer)	(Filing Date) (Anmeldedatum)	(Status) (patentiert, anhangig, aufgeben)	(Status) (patented, pending, abandoned)

Ich erkläre hiermit, dass alle von mir in der vorliegenden Erklärung gemachten Angaben nach meinem besten Wissen und Gewissen der vollen Wahrheit entsprechen, und dass ich diese eidesstattliche Erklärung in Kenntnis dessen abgebe, dass wissentlich und vorsätzlich falsche Angaben gemäss Paragraph 1001, Absatz 18 der Zivilprozessordnung der Vereinigten Staaten von Amerika mit Geldstrafe belegt und/oder Gefängnis bestraft werden koennen, und dass derartig wissentlich und vorsätzlich falsche Angaben die Gültigkeit der vorliegenden Patentanmeldung oder eines darauf erteilten Patentes gefährden können.

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true, and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

German Language Declaration

VERTRETUNGSVOLLMACHT: Als benannter Erfinder beauftrage ich hiermit den nachstehend benannten Patentanwalt (oder die nachstehend benannten Patentanwälte) und/oder Patent-Agenten mit der Verfolgung der vorliegenden Patentanmeldung sowie mit der Abwicklung aller damit verbundenen Geschäfte vor dem Patent- und Warenzeichenamt: *(Name und Registrationsnummer anführen)*

POWER OF ATTORNEY: As a named inventor, I hereby appoint the following attorney(s) and/or agent(s) to prosecute this application and transact all business in the Patent and Trademark Office connected therewith. *(list name and registration number)*

And I hereby appoint

Messrs. John D. Simpson (Registration No. 19,842) Lewis T. Steadman (17,074), William C. Stueber (16,453), P. Phillips Connor (19,259), Dennis A. Gross (24,410), Marvin Moody (16,549), Steven H. Noll (28,982), Brett A. Valiquet (27,841), Thomas I. Ross (29,275), Kevin W. Guynn (29,927), Edward A. Lehmann (22,312), James D. Hobart (24,149), Robert M. Barrett (30,142), James Van Santen (16,584), J. Arthur Gross (13,615), Richard J. Schwarz (13,472) and Melvin A. Robinson (31,870), David R. Metzger (32,919), John R. Garrett (27,888) all members of the firm of Hill, Steadman & Simpson, A Professional Corporation.

Telefongespräche bitte richten an: *(Name und Telefonnummer)*

Direct Telephone Calls to: *(name and telephone number)*

312/876-0200
Ext. ________________

Postanschrift:

Send Correspondence to:

HILL, STEADMAN & SIMPSON
A Professional Corporation
85th Floor Sears Tower, Chicago, Illinois 60606

Voller Name des einzigen oder ursprünglichen Erfinders: **NEUNEIER, Ralf**	*Full name of sole or first inventor:*
Unterschrift des Erfinders Datum 6.9.99	*Inventor's signature* Date
Wohnsitz D-81667 München, Germany DEX	*Residence*
Staatsangehörigkeit	*Citizenship*
Postanschrift Gravelottestr. 3	Post Office Addess
D-81667 München Bundesrepublik Deutschland	
Voller Name des zweiten Miterfinders (falls zutreffend): **MIHATSCH, Oliver**	*Full name of second joint inventor, if any:*
Unterschrift des Erfinders Datum O. Mihatsch 6.9.99	Second Inventor's signature Date
Wohnsitz D-80634 München, Germany DEX	Residence
Staatsangehörigkeit Bundesrepublik Deutschland	Citizenship
Postanschrift Schulstr. 31	Post Office Address
D-80634 München Bundesrepublik Deutschland	

(Bitte entsprechende Informationen und Unterschriften im Falle von dritten und weiteren Miterfindern angeben).

(Supply similar information and signature for third and subsequent joint inventors).